LANDLORD DATA METHODOLOGY

Tech Tools for Justice

White Paper

Author: Dylan Pederson

January 20th, 2025

Table of Contents

ABSTRACT

Property taxpayer record data in the United States is generally publicly available, however there are little if any data quality standards enforced for property owners submitting this data. Many landlords misspell their name and mailing address, both in the property taxpayer data and corporate and LLC data. These ubiquitous clerical errors have real consequences for city officials and researchers attempting to identify networks of property ownership. This is referred to as "ownership obscurity" in the literature. This report proposes a methodology for cleaning and matching taxpayer records that combines string matching techniques with network graph generation to identify and isolate ownership networks. Results are combined with analog research on the businesses and organizations associated with the most frequently occurring taxpayer addresses in the dataset after validation. This allows for the creation of a data model that accounts for the different ownership structures and associations (e.g., property managers vs. 'true owners'). It concludes with limitations of the methodology, room for improvement and possible paths forward for evolving the strategies used to uncover property ownership.

I. INTRODUCTION

The use of shell companies, trusts and other legal instruments allows landlords to conceal their identity, thereby obscuring the true extent of their holdings. Property ownership by wealth management corporations further complicates the process of discovering who owns what in American cities.

The academic literature refers to this as "ownership obscurity". It has real consequences not only for tenants who require their landlord's name and mailing address to send legal correspondence and seek justice in the courts, but also for local officials attempting to crack down on criminal landlords with exorbitant code violations and a track record of harassment and retaliation against whistleblowers and tenant organizers.

The basic premise of this study, of Tech Tools for Justice (TT4J) and the Landlord Mapper initiative broadly, is to reframe the conversation about power and influence in American cities towards a material understanding of urban land ownership by leveraging open data to promote transparency in rental markets. With this report, TT4J hopes to contribute to a growing body of research aimed at uncovering land ownership networks with the goal of helping tenants and local officials identify landlords and hold them to account.

As the housing crisis worsens, a growing class of perpetual renters are staring down a lifetime of dealing with landlords. This trend further necessitates greater transparency into who these landlords are, what they own, the conditions of their buildings, the treatment of tenants, and their financials.

This paper begins with a brief review of the literature and past projects which have tackled the challenge of uncovering property ownership in various cities across the US. Next, it describes the data, followed by a detailed breakdown of the methodology used to produce the Chicago landlord dataset. Then, it reviews the results and outlines the limitations of the methodology and the areas in which it can be improved.

II. TERMS & DEFINITIONS

Ownership Obscurity: Refers to the deliberate use of complex legal structures like shell companies, trusts, and registered agent services by landlords to hide their identities and the true extent of their holdings.

Shell Company: A business entity that exists only on paper, with no active business operations or significant assets. These companies are typically created to serve as a vehicle for business transactions like property ownership, often functioning as an intermediary between the true owner and their assets.

Trust: A legal arrangement where property ownership is transferred to a trustee who manages it on behalf of the beneficiary owners. The trust itself becomes the legal owner of record, while the true owners maintain control and receive benefits as beneficiaries.

Landlord Network: An association of property taxpayer records based on matching names and addresses. Networks can be used as proxies for approximating true ownership.

Landlord Entity: Any organization associated with a property taxpayer mailing address. It cannot be assumed that Landlord Entities are the "true owners" of their associated properties.

Landlord Organization: Property management companies, real estate development companies, wealth & asset management companies, and realty companies. Although it cannot be said that these organizations are the "true owners" of the properties associated with them in the taxpayer data, they are distinct from Landlord Entities in that they can usually be held accountable for living conditions in their buildings and treatment of tenants.

III. THE PROBLEM: OWNERSHIP OBSCURITY

Landlords wield their immense financial and institutional power in the United States to establish complex webs of shell companies and trusts to obscure their property holdings. For example, a renter may search for their property on the county assessor's website only to find that the name of their landlord is "123 OAK STREET LLC", or "CHICAGO LAND TITLE TRUST COMPANY 827491637". The associated mailing addresses also are often registered agent companies, office buildings without suite numbers, lockbox or virtual mail services, or the property itself. This creates massive barriers for researchers to uncover networks of property ownerships.

The problem of ownership obscurity is exacerbated by unenforced or non-existent standards for submitting property tax information. Local governments allow landlords to submit taxpayer information with incorrect spelling of names and unvalidated addresses. Local governments also usually fail to require landlords to register themselves and their properties with the city. Ownership obscurity therefore is fundamentally a problem of municipal data administration.

The consequences of ownership obscurity are experienced most directly by tenants. Many landlords do not provide their tenants with information about who they are. Even in municipalities which have passed ordinances requiring landlords to disclose this information to their tenants, many simply ignore the ordinance to keep their tenants in the dark, intentionally or otherwise. This information is necessary not only for legal correspondence, but also to blow the whistle on criminal landlords and organize tenant associations, particularly in situations where a single landlord owns many different buildings. Without the ability to send legal correspondence to their landlords, tenants have few if any avenues to seek justice.

Ownership obscurity also has implications for local officials attempting to regulate or otherwise reign in problematic landlords. For example, assume Building A has an exorbitant amount of code violations and the city decides to sanction the landlord until the code violations are rectified. This landlord owns Building A in a shell company, whose manager/member and agent are both registered agents. The landlord might own hundreds of other properties throughout the city that also have exorbitant code violations, however city officials cannot know this without laborious manual research using the county assessor and corporate and LLC databases provided by the state or third parties. And even with the tools available, it's not

guaranteed that the true owner can be discovered. Without this information, local officials are unable to hold big landlords accountable.

At a broader societal level, ownership obscurity allows powerful actors to operate in the shadows free from public scrutiny. It is essential for any free, open, democratic society that the general public can access information about who holds power over them, their communities and their cities. Ownership obscurity therefore directly undermines the principles of democratic participation that supposedly underlie American institutions.

IV. PREVIOUS RESEARCH

The development of the Chicago landlord data methodology borrows heavily from two previous research efforts to combat ownership obscurity.

The first was a project spearheaded by Forrest Hangen and Daniel T. O'Brien (Northeastern University, Harvard University) to produce landlord data for Boston. The core of their methodology is the linking of taxpayer records using a "customized, probabilistic matching algorithm to link matches with slight text variations but that still have a high degree of face validity".[1] When combined with exact matches for cleaned taxpayer names and validated addresses, Hangen et al. were able to uncover networks that could not be revealed through simply matching taxpayer names and addresses.

The second was a research initiative carried out by John Johnson and Mitchel Henke (Marquette University) to obtain landlord data for Milwaukee. Their methodology focused mainly on generating connected components via network graph objects based on cleaned names and validated addresses. Johnson et al's methodology accounts for common names and addresses which should not be used in network graph generation given their ambiguous nature. For example, if "John Smith" is the taxpayer name, it should not be assumed that all John Smiths in the dataset are associated with the same individual, therefore that name should be excluded from the network graph generation. A similar situation arises with taxpayer addresses associated with registered agents, law offices, or any other organization that cannot be assumed to be the "true owner" of the property.

The Chicago landlord data methodology synthesizes different aspects from these two projects, while contributing to the body of research by proposing additional strategies that seek to enhance the depth and specificity of the data model.

[1] Hangen et al. 2022, pg. 9

V. DATA

1. Sources

The most prominent data source for the Chicago landlord dataset is the Cook County Assessor's Office website. Property taxpayer names and mailing addresses were scraped from property detail's page using the PIN numbers. Corporate and LLC data was obtained using a parser that processes raw data made available by the Office of the Illinois Secretary of State.[2] All other data related to the properties themselves are obtained from the open data portals of the City of Chicago and Cook County.

2. Limitations

There are significant limitations to the accuracy of the Chicago landlord dataset due to the poor quality of the Chicago property taxpayer data. As outlined earlier in this document, the utter lack of data quality standards and enforcement means that landlords are allowed to submit their taxpayer data filled with errors and misspellings. This is evidenced by the wide range of mistakes observed in the raw data, such as incorrect zip codes, misspelled street names and cities, and multiple different spellings of the same names. See section VIII (Limitations) of this document to learn more.

[2] All credit for obtaining the corporate and LLC data for this project goes to the ATU-CUT landlord research collective.

VI. METHODOLOGY

The Chicago landlord data methodology is a synthesis of different strategies implemented in the previous research outlined above. The following is a general description of each stage of the process from start to finish.

1. Taxpayer Name & Address Cleaning

Preliminary data cleaning for the raw taxpayer record data included correcting common misspellings, removing multiple spaces and symbols, correcting street predirectionals (i.e., "NORTH" gets changed to "N"), and standardizing common abbreviations. For example, there were 87 different variations for the Chicago Title Land Trust Company identified in the raw data. These were replaced with a single standardized name (see figure 1).

2. Identifying Trusts and their UIDs

A major challenge in attempting to link properties to landlords is the widespread use of real estate trusts by landlords. In many cases it is downright impossible to identify the true owner of a trust-held property without a court order that would force the trust-making institution to disclose such information. However, by linking properties via network graph generation it is possible associate "true owners" with some trusts (see figure 2).

To prepare the data for identifying trusts, trust-related institutions and taxpayers found in the raw taxpayer name data were identified and flagged as trust properties. Since many taxpayer names of trust-held properties include unique identifiers, these unique IDs were extracted from the taxpayer record by first standardizing the trust institution's name, then removing the standardized name and storing the remaining characters as that property's trust's unique

taxpayer_name	taxpayer_address
CHICAGO TITLE LAND TRUST COMPANY 8002376247	7017 N KEDVALE AVE, LINCOLNWOOD, IL 60712
YAKUB AND NASIMA LAKADA	7017 N KEDVALE AVE, LINCOLNWOOD, IL 60712
CHICAGO TITLE LAND TRUST COMPANY 8002376247	7017 N KEDVALE AVE, LINCOLNWOOD, IL 60712
CHICAGO TITLE LAND TRUST COMPANY 8002376247	7017 N KEDVALE AVE, LINCOLNWOOD, IL 60712
CHICAGO TITLE LAND TRUST COMPANY 8002376247	7017 N KEDVALE AVE, LINCOLNWOOD, IL 60712
CHICAGO TITLE LAND TRUST COMPANY 3644	7017 N KEDVALE AVE, LINCOLNWOOD, IL 60712
CHICAGO TITLE LAND TRUST COMPANY 8002376247	7017 N KEDVALE AVE, LINCOLNWOOD, IL 60712
YAKUB LAKADA	7017 N KEDVALE AVE, LINCOLNWOOD, IL 60712
CHICAGO TITLE LAND TRUST COMPANY 8002376247	7017 N KEDVALE AVE, LINCOLNWOOD, IL 60712
LAKADA MANAGEMENT LLC	7017 N KEDVALE AVE, LINCOLNWOOD, IL 60712

FIGURE 1: Associating trust-held properties with landlord networks using validated address matching.

```
"THE CHICAGO TRUSTNA",
"THE CHICAGO TRUST NA A",
"THE CHICAGO TRUST NA I",
"THE CHICAGO TRUST NA T",
"THE CHICAGO TRUST  NA",
"THE CHICAGO TRUST CO",
"THE CHICAGO TRUST TR D",
"THE CHICAGO TRUST TR",
"CHICAGO TRUSTNATRUSTEE",
"CHICAGO TRUST ADMIN",
"CHICAGO TRUST COMAPNY",
"CHICAGO TRUST COMPANY",
"CHICAGO TRUST CO",
"CHICAGO TRUST TRUST",
"CHICAGO TRUST KNOWN AS",
"CHICAGO TRUST AS S",
"CHICAGO TRUST",
"CHICAGO TITLE LAND TRUST COMPANY TRUST",
"CHICAGO TITLE LAND TRUST COMPANY COMPANY"
"CHICAGO TITLE LAND TRUST COMPANY SUCCSR TTEE",
"CHICAGO TITLE LAND TRUST COMPANY LAND TRUST",
"CHICAGO TITLE LAND TRUST COMPANY AND T",
"CHICAGO TITILE LAND TRUST",
"CHICAGO TITLE LAND TRUST",
"CHICAGO TITLE LAND TRUS",
"CHICAGO TITLE LAND TRU",
"CHICAGO TITLE LAND TRT",
"CHICAGO TITLE LAND TRS",
"CHICAGO TITLE LAND TR",
"CHICAGO TITLE LAND AS",
"CHICAGO TITLE LAND CO",
"CHICAGO TITLE LAND",
"CHICAGO TITLE LND TRST",
"CHIICAGO TITLE & LAND",
"CHICAGO TITLE & TRUSTE",
"CHICAGO TITLE & TRUST",
"CHICAGO TITLE & LAND T",
"CHICAGO TITLE AS TRUST",
"CHICAGO TITLE TRUST AS",
"CHICAGO TITLE  TRUST",
"CHICAGO TITLE TRUST AG",
"CHICAGO TITLE TRUSTEE",
"CHICAGO TITLE TRUST",
"CHICAGO TITLE TR",
"CHGO TITLE LAND TRUST",
"CHGO TITLE LAND TR",
"CHICAGO TITLE  LAND TR",
"CHICAGO TITLE LAN TRUS",
"CHICAGO TITLE AND",
"CHICAGO TITLE LSND TRU",
"CHICAGO TITLE",
"C T L T COMPANY",
"CHICAGOTRST",
"CHICAGO LAND TRUST COMPANY",
"CHICAGO LAND TRUST CO",
"CHICAGO LAND TRUST TRUST",
"CHICAGO LAND TRUST TRU",
"CHICAGO LAND TRUST AS",
"CHICAGO LANDTRUST",
"CHICAGO LAND TRUST",
"CT LAND TRUST",
"CT LND TRUST",
"CT T TRUST",
"CT TRUST",
"CTANTT TRUST",
"CTCTL",
"CTLTC SUCCSR TTEE",
"CTLTC TRUST NO",
"CTLTC TRUST NUMBER",
"CTLTC TRUST",
"CTLTC TRST",
"CTLTC TR",
"CTLTC NO",
"CTLTC  NO",
"CTLTC",
"CTLT TRUST",
"CTLT AS TRUSTEE",
"CTLT NO",
"CTLT COMPANY",
"CTLT CO",
"CTLT",
"CTLCT TRUST",
"CTL TR",
"CTT LAND TRUST",
"CTT TRUSTEE",
"CTT TRUST"
"CTTRUST COMPANY",
```

FIGURE 2: Different variations of “CHICAGO TITLE LAND TRUST COMPANY” identified in the raw taxpayer data.

ID. The unique ID and trust institution name was then associated with property’s taxpayer record.[3] This permits the construction of a data model that isolates trusts as a distinct domain entity.

3. Address Validation

Address validation is the core aspect of the methodology. Given the lack of data standards enforced by municipalities for property taxpayer information, many

[3] Note that this method of identifying trust-held properties has significant limitations due to ubiquitous clerical errors in property taxpayer data. See “Limitations” below.

addresses in the raw data contain errors. Common errors identified in the raw data included misspelt street names and cities, incorrect zip codes, incorrect street predirectionals (i.e., "N STATE ST" instead of "S STATE ST"), and missing unit numbers.

This project utilized Geocodio, a geocoding and address validation service, to validate the addresses in the raw taxpayer record data. Each unique raw address wass sent to the Geocodio API as a search query, which returned matching addresses in a validated and standardized format. Since many Geocodio address searches return multiple results, the results must be parsed, and the correct match identified. This parsing process involved checking equality of street numbers and zip codes, however due to the poor quality of raw address data submitted in the taxpayer records, not all addresses are able to be validated.

For this project, there were a total of 445,994 unique raw addresses to validate. 98% of these were successfully validated and standardized, also a small fraction of these validated addresses could be inaccurate since unit numbers in the raw data have so much variation that they are sometimes not picked up by Geocodio.[4]

4. Address Research

It cannot be assumed that the residence or organization associated with the property taxpayer mailing address is the "true owner" of the property. This reality is the basis for Johnson & Henke's methodology proposed in their "Milwaukee Property Ownership Network Project", which involves identifying taxpayer addresses associated with registered agents, virtual offices, law firms and other organizations that submit property tax information on behalf of the landlord.[5] Once identified, these addresses should be excluded from the network graph generation.

This methodology proposes a separate address research process to include and distinguish between property management companies, nonprofit organizations, tax service companies, and other businesses or organizations that cannot be said to be the "true owners" of the properties. To collect this data, a list of the 3000 most frequently occurring validated addresses in the dataset was generated, and each address was manually researched to identify which type of organization is associated

[4] This is a significant limitation and can be improved on in future iterations of this methodology by accounting for unit number variation in the initial data cleaning workflow. See "Limitations" at the end of this report.

[5] See Johnson & Henke's project here: https://mkepropertyownership.com/

with it. The conceptual justification for this strategy is to capture the relationship between properties and the myriad of different organizations involved in property management and ownership. Even if "true ownership" cannot be established, it is nonetheless important to establish the relationship between the organizations and properties in the data.

The following is a list of different types of businesses and organizations found after manual address research:

- Law firms
  - These range from large law firms with significant web presence and downtown offices, to small law firms without any web presence whatsoever.
  - This could perhaps indicate that there are law firms that either deal exclusively with rental property administration, or that are themselves the "true owners" of the properties.
- Banks
  - Many of these are associated with real estate trusts.
- Construction firms
- Businesses unrelated to real estate
  - Examples include manufacturing, transportation, auto body shops, import/export companies, etc.
  - It is unclear whether or not these businesses are owned by the landlord, or if they're simply the tenant of the landlord that owns the commercial property.
- Lockbox services
- Tax consulting firms
- Financial services firms
- Office buildings without suite names
- UPS stores
  - Most likely utilizing UPS lockbox/virtual mail services.
- Virtual offices
- Registered Agents

The results of this data were stored in a manually created spreadsheet with the following Boolean columns used to build the data model:

- IS_LANDLORD_ORG
  - See definition of "Landlord Organization" outlined above.

- IS_GOVT_AGENCY
  - Local, state & federal housing authorities
- IS_LAWFIRM
- IS_MISSING_SUITE
  - Used to identify addresses that should be excluded from the network graph matching process. Since missing suite numbers could result in false positives, it is better to keep them out.
- IS_FINANCIAL_SERVICES
  - Tax consulting firms, mortgage lending firms, etc.
- IS_ASSOCIATED_BUSINESS
  - Businesses unrelated to property ownership and management
- FIX_ADDRESS
  - Flags addresses that are deemed incorrect upon manual research.
- IS_VIRTUAL_OFFICE_AGENT
- IS_NONPROFIT

It is important to note that there are many more landlord entities that were not included in this study due to the inherent limitations of manual address research. Expanding manual research efforts in the future could greatly improve the depth and accuracy of the data model by identifying additional addresses that should or should not be included in the network graph generation.

It must also be noted that the process of manual address research is subject to error. Many addresses associations with businesses, particularly those located in office parks, high-rises or otherwise ambiguous locations, and are subject to change without notice. Furthermore, many addresses are of questionable integrity, in that they point to unmarked buildings or buildings whose use or associated entities are unclear.

5. Previous Research & New Contributions

This methodology attempts to synthesize two previous methodologies developed by Hangen et al and Johnson et al, while both tweaking existing strategies and contributing to the body of research by adding new elements of the data workflow. It

seeks to go a step further by building a data model that attempts to capture the complexity of property ownership legal structures.[6]

Consider the case of property management companies, which cannot be said to be the "true owners" of properties they rent, since they could be simply providing property maintenance and administration service for other landlords. This is the basis of the distinction between "Landlord Organization" and "Landlord Entity": although they're not the "true owners", it is still important to capture their relationship with the properties. In making this distinction, the data model approximates a representation of the domain by specifying "true owners" (Landlord Networks) vs. organizations that may or may not be the true owners but that ***can*** be held accountable for the conditions of the buildings and treatment of tenants (Landlord Organizations) vs. organizations that may or may not be the true owners but ***cannot*** be held accountable.[7]

The core of this methodology is the network graph. Nodes and edges are created that link taxpayer names to mailing addresses, which produces connected components (see figures 4 & 5). Corporate and LLC names and addresses are also passed into the graph, adding to the connected components created by the taxpayer records.

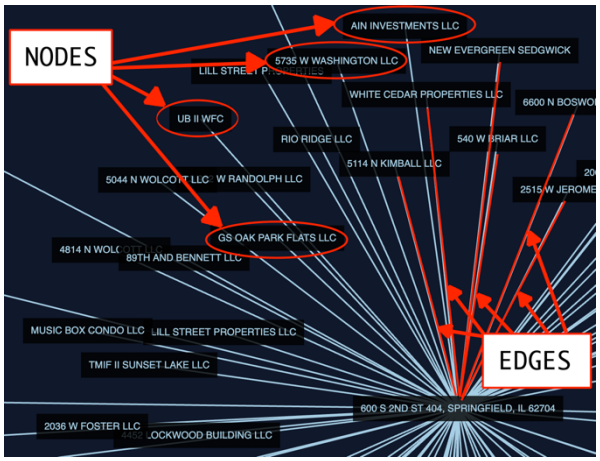

FIGURE 4: Nodes and edges of a network graph

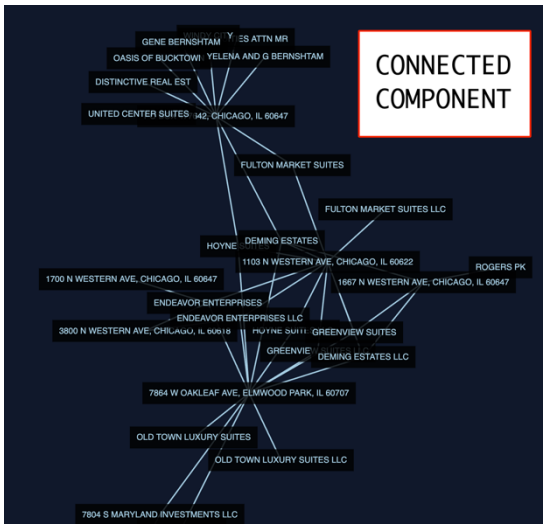

FIGURE 5: Connected component, constructed out of nodes and edges

[6] A deeper dive into better understanding of these legal structures will allow for the refinement of the data model with more accurate domain representation, and should be considered in future iterations of this methodology. See "Limitations" below.

[7] Classifying these organizations is an open and ongoing question, and should be considered in future iterations of this methodology. See "Limitations" below.

String matching taxpayer records is also a core aspect of the methodology. Matches are assigned a standardized name which is passed into the network graph as an additional node associated with the taxpayer names and addresses via edge construction. This allows for the association of properties whose taxpayer names and addresses might not match well enough to be connected to the same component.

6. Rental Property Subset

The city of Chicago maintains 182 unique building class codes, 40 of which are classified as "Residential" and "Multi-Family". Of those, 18 were chosen to subset the property scrape for rental-only properties. See Appendix for these class codes and their descriptions.

Addresses were validated ***before*** subsetting the property dataset for rental-only properties. The purpose of this approach is to be able to include properties whose class codes did ***not*** match any of the rental property codes chosen to subset the dataset, but whose validated taxpayer mailing address matches those that ***were*** included in the dataset. This allows for the inclusion of properties associated with rental residential landlord networks whose class codes were not included in the subset.

For example, consider a situation in which there are 10 properties whose validated taxpayer mailing address is "5847 N MAPLE AVE STE 101, SPRINGFIELD IL 62701". These properties' class codes do ***not*** match the rental class codes chosen to subset the data, so the initial subset does not include them. However, there are 100 properties that ***were*** included in the rental subset whose taxpayer address is also "5847 N MAPLE AVE STE 101, SPRINGFIELD IL 62701". This methodology first executes the subset, then checks all non-rental properties' taxpayer addresses against a list of unique taxpayer addresses from the rental subset and pulls in additional properties whose taxpayer addresses match those from the rental subset.

7. Setting the Parameters

Both the string matching and network graph generation processes involve setting parameters that significantly alter the outcome. These include the tolerance threshold for string-matching, the string-matching result to pass into the network graph,

whether or not to include Landlord Organization addresses in the address matching, and whether or not to include addresses that have ***not*** been researched manually.

For this study, parameter matrices were created both for the string matching and network graph generation processes, and each unique parameter combination was run on the data. The primary outcome measurements for string matching are the number of unique matches and the quantity of properties associated with the top 100 connected components with the most buildings. Sorting the resulting outputs by properties in top 100 components allows a ranking of most to least accurate Landlord Networks.

8. Indicating Inclusion of Addresses

A parameter was included in both the string matching and network graph generator to indicate whether or not certain addresses should be included in the node-edge construction. These are Landlord Organization addresses and addresses that were ***not*** included in the manual research processes outlined above in Section 4. This allowed for control over which manually researched addresses should or shouldn't be included in the network graph, or any other addresses relevant to the domain.

9. Common Names

The presence of common names in both taxpayer, corporation and LLC data presents challenges for effectively linking individual landlords to properties and organizations. Johnson et al propose checking these names against those obtained in voter registration data and filtering out names that are deemed "common" and therefore cannot be used in network graph generation. For example, if Jorge González appears hundreds of times in the voter registration data, this name will ***not*** get passed into the node and edge constructor.

It remains critical, however, that these names are used somehow in the process of uncovering networks. The string-matching process proposed by Hangen et al allow for common names to still be used to associate taxpayer records since nodes and edges are formed by N-gram measurements of the ***concatenated taxpayer names and addresses***. For example, consider the following fictitious taxpayer records:

- Enrique Jiménez -- 123 Oak St, Springfield IL 62701

- Enrique Jiménez -- 123 Oak St, Springfield IL 62701
- Enrique Jiménez -- 123 Oak St, Springfield IL 62701
- Enrique Jiménez -- 456 Maple Ave, Springfield IL 62701
- Enrique Jiménez -- 456 Maple Ave, Springfield IL 62701

It can be assumed reasonably that the first three records are the same person given their matching addresses. However, when running the network graph generator ***without*** including the string matching results, the nodes and edges linking the first three taxpayer records would never have been created.

This study therefore proposes that string matching process proposed by Hangen et al should be used ***before and in tandem with*** network graph generation as presented in Johnson et al's work. The workflow methodology proposed in this report does ***not*** exclude common names for the string matching, but it ***does*** exclude them from the network graph generation.

This study did not use voter registration data to exclude common names. Rather, a common names list was manually created via manual analysis of the most commonly appearing cleaned taxpayer names in the Chicago dataset.

10. Matching Corporations & LLCs to Taxpayer Names

Corporations and LLCs are matched to taxpayer records in two stages. The first is exact matching, which is executed by comparing strings in the cleaned taxpayer name and cleaned entity[8] name and creating associations for names that match exactly. After that, the remaining unmatched taxpayer records are run through the same string-matching configuration used to match taxpayer records described below in Section 12. Taxpayer records associated with an entity are then linked to the mailing addresses associated with that entity (office, manager/member and agent addresses of LLCs and president and secretary addresses of corporations). Upon linkage, the addresses of those entities were ised to construct nodes and edges in the network graph generation outlined in Section 13.

[8] Note that "entity" here does ***not*** refer to Landlord Entity as defined above.

11. Clean vs. Core Name

```
UNIQUE_KEYS = [
    'CIR ', 'APARTMENTS ', 'SERVICES ', 'INVESTMENTS ', 'HOLDINGS ',
    'LN ', 'COMPANY ', 'AUTHORITY ', 'INC ', 'FORECLOSURE ',
    'ESTABLISHED ', 'CONDO TRUST ', 'COOPERATIVE ', 'PARTNERS ', 'CR ',
    'PARTNERSHIP ', 'GROUP ', 'ASSOCIATION ', 'TRUSTEES ', 'TRUST ',
    'PROPERTIES ', 'MANAGEMENT ', 'SQUARE ', 'MANAGERS ', 'EXCHANGE ',
    'REAL ESTATE ', 'DEVELOPMENT ', 'REDEVELOPMENT ', 'MORTGAGE ',
    'RESIDENTIAL ', 'REALTY TRUST ', 'CORPORATION ', 'LIMITED ', 'LLC ',
    'ORGANIZATION ', 'REALTY ', 'PRT ', 'VENTURE ', 'RENTAL ', 'UNION ',
    'CONDO '
]
```

FIGURE 6: Unique keys used to generate CORE_NAME (Hangen et al)

Hangen et al propose two distinct string cleaning processes to increase exact matches by removing common keywords that if removed would match taxpayer names that wouldn't have been matched after the first round of cleaning. Taxpayer and entity names in the CLEAN_NAME column are those that underwent basic string cleaning without the removal of these keywords, while the CORE_NAME column contains names that had these keywords removed. This creates two different columns to run the string matching and network graph on, the latter of which in theory would result in a greater number of matches.

12. String Matching

The following parameters are used to construct the matrix for the string-matching process:

- CLEAN_NAME vs. CORE_NAME
- INCLUDE_ORGS?
  - When set to true, addresses associated with Landlord Organizations ***are*** included in the string matching.
- INCLUDE_UNRESEARCHED?
  - When set to true, addresses that were ***not*** manually researched are included in the string matching.
- MATCH_THRESHHOLD
  - Match confidence threshold specified when running string matching algorithm

The string match algorithm in this methodology utilizes a Term Frequency-Inverse Document Frequency matrix with a 3-gram cosine similarity to measure string similarity of cleaned taxpayer names and addresses concatenated.[9] It measures the similarity of each concatenated taxpayer record, and matches that are above the

[9] Borrowed from Hangen et al's landlord data methodology. See "Linking Landlords" (2022).

specified threshold are assigned a standardized string match ID that is a concatenation of the most commonly appearing distinct taxpayer names.

For example, consider these taxpayer records for Chicago landlord George Triff:

- GEORGE TRIFF -- PO BOX 173, OAK FOREST, IL 60452
- GEORGE TRIFF -- PO BOX 17304, OAK FOREST, IL 60617
- GEROGE TRIFF -- PO BOX 173, OAK FOREST, IL 60452
- GEORGE TRIGG -- PO BOX 173, OAK FOREST, IL 60452
- GEORGE TRIFF -- PO BOX 173, OAK FOREST, IL 60452
- GEORE TRIFF -- PO BOX 173, OAK FOREST, IL 60452

At a certain match threshold all of these taxpayer records would be associated based on the string similarity score produced by the N-gram. Each of these records would then be assigned a component name, being the most three commonly appearing names among all names associated after the string matching:

- GEORGE TRIFF -- GEORE TRIFF -- GEORGE TRIGG

After running the string matching with the different parameters in place, the NetworkX python package was used to create connected components by creating nodes and edges with taxpayer names and addresses and the component names assigned to properties after the string matching.

13. Network Graph (Taxpayer Records & Corporate/LLC Data)

After running the property tax record data through the string-matching process, the dataset is ready for the main network graph generation. Properties associated with networks derived from the string matching are associated with a component name, which will be used as a parameter set in the network graph generation.

The following are parameters used in the network graph generation:

- CLEAN_NAME vs. CORE_NAME
- INCLUDE_ORGS?
  - Same as string matching parameter.
- INCLUDE_UNRESEARCHED?
  - Same as string matching parameter.

- STRING_MATCH_RESULT
  - Specifies the string-matching result column.
  - The parameter matrix returned 16 string matching results, stored in individual data columns, meaning a specific string-matching column to use in the network graph generation must be specified.

Similar to the string matching, properties were given a unique name for the associated connected component, consisting of the three most commonly appearing taxpayer names among all taxpayer records associated with a connected component, concatenated with " -- ".

To increase address matches in the network graph, secondary address number prefixes were removed. For example, consider the following three addresses:

- 11705 S State St Unit 201, Chicago IL
- 11705 S State St Ste 201, Chicago IL
- 11705 S State St #201, Chicago IL

These would ***not*** be pulled into the same component because they are not identical. However, upon removing the prefixes they ***will*** be pulled in:

- 11705 S State St 201, Chicago IL
- 11705 S State St 201, Chicago IL
- 11705 S State St 201, Chicago IL

VII. RESULTS

The *Cook County Address Points* dataset provided by the Cook County Open Data Portal contains the PIN numbers for all properties associated with unique street addresses in Cook County. These total to 1,165,958 properties as of April 8th, 2024, the most recent date the dataset was updated at the time of this study. These properties were subsetted for the city of Chicago using their associated addresses' zip codes, totaling to 469,413 properties.

Building class codes were used to further subset this data to include only rental properties, reducing the total property count for this study to 129,007. After checking validated taxpayer addresses of this rental property subset against all Chicago properties, an additional 35,837 properties that were ***not*** initially included in the rental subset shared validated taxpayer addresses with those in the initial rental subset. Properties associated with MTO hotline data[10] were also pulled into the subset, bringing the total number of rental properties processed by the workflow to 166,165.

The total number of unique taxpayer addresses in the Chicago property dataset came out to 424,041, of which 416,568 (or 98.2%) were successfully validated. To be validated, an address must successfully return results after sending the raw address as a search query to the Geocodio API, and those results must be parsed and filtered to identify the correct one to use in the workflow. 7473 addresses remain unvalidated.

1. String Matching

Across the board, the most consequential parameter was whether or not ***unresearched addresses*** were included in the string matching. Considering only around 3000 of the 402,927 unique validated addresses in the dataset were manually researched, it is to be expected that their inclusion would significantly increase the number of matched taxpayer records.

[10] Tech Tools for Justice (TT4J) has partnered with the Metropolitan Tenants Organization (MTO), a Chicago-based non-profit that helps tenants organize unions, particularly in government-subsidized buildings and Single-Room Occupancy (SRO) buildings. MTO maintains a hotline which tenants can call to report problems in their buildings, and has agreed to share this data with TT4J to incorporate into the data model and display in the Landlord Mapper web tool.

The inclusion of Landlord Organization addresses, and the reduction of the match threshold also appeared to correlate with more matches. A slight correlation can be observed, although much less than including unresearched addresses.

Clean name vs. core name did not appear to correlate with more or less matches. The ordered results show a perfect 1:1 alternation between clean and core name suggesting it has the lowest impact of all the parameters. See the Appendix Section II for the matrix results data.

The most conservative estimate from the parameter matrix results identified 774 unique matches, with a total top 100 building count of 1669. The most liberal estimate yielded 7371 matches with 5766 top 100 buildings (see figure 8).

The following string match results were chosen to include as matrix parameters for the network graph generation outlined above in section 13 of *Methodology*, in order

Methodology Calculation	Taxpayer Name	Include Orgs	Include Unresearched	Match Threshold
STRING_MATCH_NAME_1	CLEAN	FALSE	FALSE	0.85
STRING_MATCH_NAME_5	CLEAN	TRUE	FALSE	0.85
STRING_MATCH_NAME_3	CLEAN	FALSE	TRUE	0.85
STRING_MATCH_NAME_4	CLEAN	FALSE	TRUE	0.8

FIGURE 7: Parameter definitions for each string match result used in network graph generation

Metric	Max	Min	StDev (Sample)	StDev (Population)	Variance (Sample)	Variance (Population)
Match Count	7371	774	2783	2695	7745019	7260956
Top 100 Building Count	5677	1669	1451	1405	2105131	1973560
Matched Properties	29285	4358	9735	9426	94774680	88851262
Percent Properties Matched	7371	774	2783	2695	7745019	7260956

FIGURE 8: String Matching Matrix Outcome Indicators and statistics

from least to most liberal. They were chosen as their output roughly corresponds to breaks in the distribution of matches and buildings.

2. Network Graph

The network graph results differ substantially from those of the string matching in that the observed correlations between parameters and outcome metrics change depending on how the results are sorted. When changing the sort column of matrix results from top 100 building count to number of unique networks, different patterns and correlations of individual parameters can be observed.

Unique network count is a proxy for degree of concentration of property ownership, as fewer calculated networks means that a greater number of validated addresses were pulled into the connected component, resulting in fewer outputted unique networks. The most impactful parameter for the network count sort results is the again the inclusion of unresearched addresses. There is also a slight correlation with the inclusion of Landlord Organization addresses.

Sorting by number of 100 buildings switches up the parameter outcomes, with the inclusion of Landlord Organizations correlating starkly with number of buildings while inclusion of unresearched does not appear to have any correlation. As with the string matching results, the difference between clean name and core name did not appear to correlate, regardless of sort column.

An additional parameter used in the network graph generation matrix was the entity clean and core name. Not to be confused with Landlord Entity defined above, entity

Metric	Max	Min	StDev (Sample)	StDev (Population)	Variance
Network Count	122221	94815	10353	10272	107190858
Top 100 Building Count	13306	5800	2197	2179	4824879
Buildings w/o Network	9590	2984	2995	2971	8967525

FIGURE 9: Network Graph Matrix Outcome, Indicators and statistics

in this context refers to either a corporation or LLC that was matched with the property above in Section 10 of the methodology. The same general principle between clean name and core name applies, however instead of just including a core name column for taxpayer name, the network graph also includes it for entity name.

Columns were also included for whether or not unresearched addresses and Landlord Organizations were included in the ***string-matching result*** used as a parameter. For example, if unresearched addresses were included in the string match parameter matrix, STRING_MATCHED_NAME_3 would be 'TRUE' in the "Include unresearched string" column.

When sorted by top 100, the lowest output yielded 5800, while the highest yielded 13,306. Unique network counts showed less variation, with 122,221 at the high end and 94,815 at the low end. See Appendix for sorted tables and figure 9 for summary statistics and measures of distribution.

Out of the 64 outcomes resulting from the network parameter matrix, six were selected to include in the final data model. Similar to choosing thew string matching results for the network graph, six roughly evenly distributed break points were chosen within the Top 100 Buildings sort. These six results, ordered from least to most liberal, comprise Methodologies 1-6. See figure 10 for the parameter definitions used to generate the methodologies appearing on the Landlord Mapper web tool.

Methodology Calculation	Taxpayer Name	Entity Name	Include Orgs	Include Orgs (String)	Include Unresearched	Include Unresearched (String)	String Match Result
Methodology 1	CLEAN	CLEAN	FALSE	FALSE	FALSE	FALSE	STRING_MATCH_NAME_1
Methodology 2	CLEAN	CORE	FALSE	FALSE	FALSE	TRUE	STRING_MATCH_NAME_4
Methodology 3	CORE	CLEAN	FALSE	FALSE	TRUE	TRUE	STRING_MATCH_NAME_4
Methodology 4	CLEAN	CLEAN	TRUE	FALSE	FALSE	TRUE	STRING_MATCH_NAME_3
Methodology 5	CLEAN	CLEAN	TRUE	FALSE	TRUE	TRUE	STRING_MATCH_NAME_4
Methodology 6	CORE	CLEAN	TRUE	FALSE	TRUE	TRUE	STRING_MATCH_NAME_4

FIGURE 10: Network graph parameter matrix definitions for select results used to populate the Landlord Mapper database

VIII. LIMITATIONS & FUTURE RESEARCH

The poor quality of publicly available property taxpayer data has been established as a major limitation to this kind of study. Many local governments fail to enforce data quality standards for property taxpayer information, which combined with outdated taxpayer records means that the landlord networks calculated by the data workflow are subject to error, mostly false positives. There is little to be done about this limitation, as it is up to local governments to properly enforce such standards.

Address validation is another limitation that could lead to the generation of false positives, or more likely it would cause taxpayer records that ***should*** be included in a given network to not be. This again is related to the poor quality of taxpayer information submitted by property owners, as misspellings in the street names, variation in street type suffixes and secondary number prefixes, incorrectly specified zip codes, etc. A significant area of improvement for this study could be a more rigorous approach to address cleaning pre-validation, and a more granular approach to parsing and filtering validated addresses returned from the search Geocodio query.

Beyond data source integrity, another major limitation is the classification of Landlord Entities and Organizations. While the definitions proposed in this report are based on real-life domain relationships, the specificity and intricacy of these legal entities and the individuals associated with them is more complex and requires a more granular approach. Future studies should more carefully model the domain to approximate relationships between entities that most precisely represents real-life legal structures and relationships that facilitate ownership obscurity.

The process of matching corporation and LLC names to taxpayer records presents yet another limitation to the outcome of the Landlord Network calculations. Many corporations and LLCs in the taxpayer data remain unmatched to those in the Illinois Secretary of State's corporation and LLC datasets. This is due to the fact that even after running exact matching ***and*** string matching on the cleaned names, some of them do not get picked up. Future iterations of this type of research should refine matching strategies to pick up on ones left out by the process outlined in this report.

IX. CONCLUSION

The present study proposes a methodology for obtaining owner-linked property datasets from taxpayer data that utilizes string matching and network graph generation to associated properties to their “true owners”. It borrows strategies developed by researchers previously while suggesting adjustments and new processes that serve the aim of modeling the domain at the database level. It also hopes to inform future studies that seek to produce such datasets with greater depth and domain specificity.

X. APPENDIX

I. Building class codes used for rental subset

MAJOR CLASS 2: Residential

- 211
  - Apartment building with 2 to 6 units, any age
- 219
  - A residential building licensed as a Bed & Breakfast by the municipality, County of Cook, or registered as a Bed & Breakfast with the State of Illinois under 50 ILCS 820/1 et seq., with six or fewer rentable units and where none of the units are owner occupied and no homeowner's exemption is allowed under the Property Tax Code
- 225
  - Single-room occupancy ("SRO") rental building

MAJOR CLASS 3: Multi-Family

- 313
  - Two-or-three story, building, seven or more units
- 314
  - Two-or-three-story, non-fireproof building with corridor apartment or California type apartments, no corridors, exterior entrance
- 315
  - Two-or-three story, non-fireproof corridor apartments or California type apartments, interior entrance
- 318
  - Mixed-use commercial/residential building with apartment and commercial area totaling 7 units or more or between 20,000 to 99,999 square feet of building area, with the commercial component of the property consisting of no more than 35% of the total rentable square footage
- 391
  - Apartment building over three stories, seven or more units
- 396
  - Rented modern row houses, seven or more units in a single development or one or more contiguous parcels in common ownership

- 397
  - Special rental structure
- 399
  - Rental condominium

MAJOR CLASS 9: Class 3 Multi-Family Residential Real Estate Incentive

- 913
  - Two-or-three-story apartment building, seven or more units
- 914
  - Two-or-three-story non-fireproof court and corridor apartments or California type apartments, no corridors, exterior entrance
- 915
  - Two-or-three-story non-fireproof corridor apartments, or California type apartments, interior entrance
- 918
  - Mixed-use commercial/residential building with apartments and commercial area where the commercial area is granted an incentive use
- 959
  - Rental condominium unit
- 991
  - Apartment buildings over three stories
- 996
  - Rented modern row houses, seven or more units in a single development or one or more contiguous parcels in common ownership

II. String matching results, ordered by Top 100 Building Count

taxpayer_col	include_orgs?	include_unresearched?	match_threshold	match_name	unique_tax_records_clean	unique_tax_records_core	matched_properties	unique_matches	percent_matched	top_100_bldg_count
CORE_NAME	FALSE	FALSE	0.85	STRING_MATCHED_NAME_9	139247	138757	4358	774	0.56%	1669
CLEAN_NAME	FALSE	FALSE	0.85	STRING_MATCHED_NAME_1	139247	138757	4614	813	0.58%	1743
CORE_NAME	FALSE	FALSE	0.8	STRING_MATCHED_NAME_10	139247	138757	5593	962	0.69%	1945
CLEAN_NAME	FALSE	FALSE	0.8	STRING_MATCHED_NAME_2	139247	138757	5753	1016	0.73%	1904
CORE_NAME	TRUE	FALSE	0.85	STRING_MATCHED_NAME_13	139247	138757	6173	987	0.71%	2499
CLEAN_NAME	TRUE	FALSE	0.85	STRING_MATCHED_NAME_5	139247	138757	6731	1055	0.76%	2662
CORE_NAME	TRUE	FALSE	0.8	STRING_MATCHED_NAME_14	139247	138757	7969	1263	0.91%	2841
CLEAN_NAME	TRUE	FALSE	0.8	STRING_MATCHED_NAME_6	139247	138757	8531	1351	0.97%	2933
CORE_NAME	FALSE	TRUE	0.85	STRING_MATCHED_NAME_11	139247	138757	19782	5236	3.77%	4132
CLEAN_NAME	FALSE	TRUE	0.85	STRING_MATCHED_NAME_3	139247	138757	20553	5370	3.86%	4379
CORE_NAME	TRUE	TRUE	0.85	STRING_MATCHED_NAME_15	139247	138757	21692	5468	3.94%	4764
CLEAN_NAME	TRUE	TRUE	0.85	STRING_MATCHED_NAME_7	139247	138757	23020	5688	4.09%	5113
CORE_NAME	FALSE	TRUE	0.8	STRING_MATCHED_NAME_12	139247	138757	25697	6859	4.94%	4779
CLEAN_NAME	FALSE	TRUE	0.8	STRING_MATCHED_NAME_4	139247	138757	26321	7009	5.03%	4837
CORE_NAME	TRUE	TRUE	0.8	STRING_MATCHED_NAME_16	139247	138757	28444	7189	5.18%	5576
CLEAN_NAME	TRUE	TRUE	0.8	STRING_MATCHED_NAME_8	139247	138757	29285	7371	5.29%	5677

III. Network graph results (1-32), ordered by Top 100 Building Count

taxpayer_col	entity_col	include_orgs?	include_orgs_string?	include_unresearched?	include_unresearched_string?	string_match_name	network_name	bldg_count_no_ntwk	unique_ntwk_count	top_100_bldg_count
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_1	3717	122221	5800
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_17	3717	122209	5810
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_2	3709	121788	5976
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_18	3709	121776	5986
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_33	3656	121420	6263
CORE_NAME	ENTITY_CORE_NAME	FALSE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_49	3656	121420	6263
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_34	3656	121047	6420
CORE_NAME	ENTITY_CORE_NAME	FALSE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_50	3656	121047	6420
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_5	9590	99904	6806
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_21	9587	99855	6900
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_6	9582	99480	6979
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_3	3254	117100	7022
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_19	3254	117089	7031
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_22	9579	99433	7071
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_4	3058	115381	7286
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_20	3058	115370	7295
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_35	3203	116487	7448
CORE_NAME	ENTITY_CORE_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_51	3203	116487	7448
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_37	9488	99344	7557
CORE_NAME	ENTITY_CORE_NAME	FALSE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_53	9490	99347	7557
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_36	3006	114794	7706
CORE_NAME	ENTITY_CORE_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_52	3006	114794	7706
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_38	9488	98976	7718
CORE_NAME	ENTITY_CORE_NAME	FALSE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_54	9490	98979	7718
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_7	9284	98674	7977
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_23	9281	98625	8069
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_8	9095	98400	8204
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_24	9092	98351	8297
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_39	9183	98146	8687
CORE_NAME	ENTITY_CORE_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_55	9185	98149	8687
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_40	8997	97880	8912
CORE_NAME	ENTITY_CORE_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_56	8999	97883	8912

IV. Network graph results (33-64), ordered by Top 100 Building Count

taxpayer_col	entity_col	include_orgs?	include_orgs_string?	include_unresearched?	include_unresearched_string?	string_match_name	network_name	bldg_count_no_ntwk	unique_ntwk_count	top_100_bldg_count
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_9	3657	118551	9304
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_10	3657	118548	9306
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_25	3657	118535	9417
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_26	3657	118532	9419
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_41	3634	117856	10138
CORE_NAME	ENTITY_CORE_NAME	TRUE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_57	3634	117856	10138
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_42	3634	117853	10140
CORE_NAME	ENTITY_CORE_NAME	TRUE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_58	3634	117853	10140
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_11	3194	113693	10354
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_27	3194	113678	10471
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_12	2998	112062	10528
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_28	2998	112047	10645
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_13	9510	96436	10658
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_14	9510	96436	10658
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_29	9506	96382	10850
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_30	9506	96382	10850
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_43	3181	113162	11208
CORE_NAME	ENTITY_CORE_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_59	3181	113162	11208
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_44	2984	111554	11378
CORE_NAME	ENTITY_CORE_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_60	2984	111554	11378
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_15	9203	95449	11722
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_16	9014	95261	11870
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_31	9199	95395	11916
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_32	9010	95207	12064
CORE_NAME	ENTITY_CORE_NAME	TRUE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_61	9448	95981	12071
CORE_NAME	ENTITY_CORE_NAME	TRUE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_62	9448	95981	12071
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_45	9446	95978	12088
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_46	9446	95978	12088
CORE_NAME	ENTITY_CORE_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_63	9142	95001	13138
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_47	9140	94998	13155
CORE_NAME	ENTITY_CORE_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_64	8956	94818	13289
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_48	8954	94815	13306

V. Network graph results (1-32), ordered by Unique Network Count

taxpayer_col	entity_col	include_orgs?	include_orgs_string?	include_unresearched?	include_unresearched_string?	string_match_name	network_name	bldg_count_no_ntwk	unique_ntwk_count	top_100_bldg_count
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_1	3717	122221	5800
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_1	3717	122209	5810
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_2	3709	121788	5976
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_1	3709	121776	5986
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_3	3656	121420	6263
CORE_NAME	ENTITY_CORE_NAME	FALSE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_4	3656	121420	6263
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_3	3656	121047	6420
CORE_NAME	ENTITY_CORE_NAME	FALSE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_5	3656	121047	6420
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_9	3657	118551	9304
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_1	3657	118548	9306
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_2	3657	118535	9417
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_2	3657	118532	9419
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_4	3634	117856	10138
CORE_NAME	ENTITY_CORE_NAME	TRUE	FALSE	FALSE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_5	3634	117856	10138
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_4	3634	117853	10140
CORE_NAME	ENTITY_CORE_NAME	TRUE	TRUE	FALSE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_5	3634	117853	10140
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_3	3254	117100	7022
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_1	3254	117089	7031
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_3	3203	116487	7448
CORE_NAME	ENTITY_CORE_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_5	3203	116487	7448
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_4	3058	115381	7286
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_2	3058	115370	7295
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_3	3006	114794	7706
CORE_NAME	ENTITY_CORE_NAME	FALSE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_5	3006	114794	7706
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_1	3194	113693	10354
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_2	3194	113678	10471
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_4	3181	113162	11208
CORE_NAME	ENTITY_CORE_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_5	3181	113162	11208
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_1	2998	112062	10528
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_2	2998	112047	10645
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_4	2984	111554	11378
CORE_NAME	ENTITY_CORE_NAME	TRUE	FALSE	FALSE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_6	2984	111554	11378

VI. Network graph results (33-24). ordered by Unique Network Count

taxpayer_col	entity_col	include_orgs?	include_orgs_string?	include_unresearched?	include_unresearched_string?	string_match_name	network_name	bldg_count_no_ntwk	unique_ntwk_count	top_100_bldg_count
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_5	9590	99904	6806
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_21	9587	99855	6900
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_6	9582	99480	6979
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_22	9579	99433	7071
CORE_NAME	ENTITY_CORE_NAME	FALSE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_53	9490	99347	7557
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_37	9488	99344	7557
CORE_NAME	ENTITY_CORE_NAME	FALSE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_54	9490	98979	7718
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_38	9488	98976	7718
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_7	9284	98674	7977
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_23	9281	98625	8069
CLEAN_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_8	9095	98400	8204
CLEAN_NAME	ENTITY_CORE_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_24	9092	98351	8297
CORE_NAME	ENTITY_CORE_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_55	9185	98149	8687
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_39	9183	98146	8687
CORE_NAME	ENTITY_CORE_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_56	8999	97883	8912
CORE_NAME	ENTITY_CLEAN_NAME	FALSE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_40	8997	97880	8912
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_13	9510	96436	10658
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_14	9510	96436	10658
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_29	9506	96382	10850
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_30	9506	96382	10850
CORE_NAME	ENTITY_CORE_NAME	TRUE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_61	9448	95981	12071
CORE_NAME	ENTITY_CORE_NAME	TRUE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_62	9448	95981	12071
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	TRUE	FALSE	STRING_MATCHED_NAME_1	CHI_NETWORK_45	9446	95978	12088
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	TRUE	TRUE	FALSE	STRING_MATCHED_NAME_5	CHI_NETWORK_46	9446	95978	12088
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_15	9203	95449	11722
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_31	9199	95395	11916
CLEAN_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_16	9014	95261	11870
CLEAN_NAME	ENTITY_CORE_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_32	9010	95207	12064
CORE_NAME	ENTITY_CORE_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_63	9142	95001	13138
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_3	CHI_NETWORK_47	9140	94998	13155
CORE_NAME	ENTITY_CORE_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_64	8956	94818	13289
CORE_NAME	ENTITY_CLEAN_NAME	TRUE	FALSE	TRUE	TRUE	STRING_MATCHED_NAME_4	CHI_NETWORK_48	8954	94815	13306