# LINKING ENTITY TAX RETURNS AND WAGE FILINGS

Prepared by the Staff
of the
JOINT COMMITTEE ON TAXATION

April 5, 2022
JCX-5-22

# CONTENTS

Page

# INTRODUCTION

One of the responsibilities of the staff of the Joint Committee on Taxation ("Joint Committee staff") is to provide Congress with estimates of the budgetary impacts of proposed tax legislation. This starts with an economic analysis of the proposed legislation.

The Joint Committee staff often leverages multiple sources of data as part of an analysis. The core of most analyses begins with confidential databases of individual and entity-level tax returns and associated filings provided by the Internal Revenue Service ("IRS"). While these data may contain related information and are generated from similar reporting procedures, it is often difficult to match data from different sources. This is the case for employer tax returns (*e.g.*, Form 1120, *U.S. Corporation Income Tax Return*) and individual wage filings (*i.e.*, Form W-2, *Wage and Tax Statement*).

Wages paid to employees by for-profit entities are generally reported as line entries on the income tax return of the employer and on individual wage filings that the employer is required to provide employees and the IRS. Business entities are identified by employer identification numbers ("EINs") on entity tax returns and individual wage filings. However, some business entities use multiple EINs when issuing W-2s and filing other tax forms. This is because business entities are frequently organized in group structures that include a parent corporation that owns several subsidiary corporations. While the parent corporation may file a Form 1120 for the group, the subsidiary corporation may be the entity that employs workers, pays wages, and withholds tax from those. As a result, a parent-entity "EIN" may provide an incomplete picture of total wages paid and taxes withheld by a business group, because the EIN on the Form 1120 may not be the EIN on the Form W-2.

The Joint Committee staff has developed a parent-subsidiary bridge data set ("the bridge") to better link parent EINs, used to file parent-level federal income tax returns, and the EINs these entities use to file various tax forms. This bridge provides a more complete and accurate accounting of firm activities and characteristics. For example, when matching W-2 filings to a stratified-random sample of subchapter C corporations provided by the Statistics of Income division of the IRS ("SOI"), use of the bridge allows the Joint Committee staff to account for 95 percent of wage deductions claimed on Form 1120. Performing the match without using the bridge accounts for approximately 36 percent of the wages in that same sample.

The parent-subsidiary bridge is particularly useful when evaluating proposals that are conditioned on characteristics of a firm's workforce. For example, treatment under the employee retention credit contained in Public Law 116-136 — the Coronavirus Aid, Relief, and Economic Security ("CARES") Act – was based on the number of employees at a firm. Bridging parent and subsidiary entities provides a more accurate picture of the total number of employees at a firm with both parent and subsidiary entities.

This document[1] describes the process by which the Joint Committee staff constructs the bridge data set connecting parent EINs to EINs reported on individual-level wage filings, including the use of an automated string-matching process to record matches between similar but not identical firm names (the "fuzzy name matching procedure"). This document also compares the bridge to other tax data and survey data and outlines the utility of the bridge for Joint Committee revenue estimates.[2]

[1] This document may be cited as follows: Joint Committee on Taxation, *Linking Entity Tax Returns and Wage Filings* (JCX-5-22), April 5, 2022. This document is available on the Joint Committee on Taxation website at www.jct.gov.

[2] The staff of the Joint Committee on Taxation welcomes comments from interested readers who have studied modeling of the Federal individual income tax. Direct comments to Chief of Staff, Thomas A. Barthold, and Deputy Chief of Staff, Robert P. Harvey, Joint Committee on Taxation, 502 Ford House Office Building, Washington, D.C. 20515-6453.

## I. APPROACH TO MATCHING ENTITY AND INDIVIDUAL DATA

The parent entity tax returns to which the Joint Committee staff links individual wage filings include all types of corporate, partnership, and non-profit tax returns. Within these forms,[3] firms report to the IRS general information about the firm including their EIN, address, and name as well as income, deductions, and tax liability. Individual wages are reported by employers on Form W-2. This form contains taxpayer and employer identification numbers, wages paid to the individual, and other information related to employer provided benefits.

Linking these entity-level tax returns and individual-level W-2s is a challenge because many entities use multiple EINs[4] to file W-2s. As a result, a simple match between entity returns and W-2s which relies on a single entity EIN appearing on both forms leads to an incomplete match of entities to their employees. To address this deficiency, the Joint Committee staff built a parent-subsidiary bridge. The bridge is the product of several processes to link entity tax returns to individual wage filings: (1) a mechanical match using information on subsidiaries reported on Form 851 (an attachment to a parent's Form 1120) or information on qualified subchapter S subsidiaries reported on Form 8869, (2) a manual matching process using the names on entity tax returns and the payer names on Form W-2, and (3) a "fuzzy" or statistical name-matching procedure (the "fuzzy name-matching procedure"). The data used in each of these matching processes are the population of Form W-2 filings and the population of entity-level tax returns in years 1999–2019 (with subsequent years added upon availability). For example, in 2018 these data include 235 million Form W-2 filings and 12.9 million entity-level tax returns.[5]

The first step is the mechanical matching process. Corporations list the EINs and the percentage of each subsidiary which the corporation owns on Form 851, *Affiliations Schedule.* The mechanical matching process involves recording subsidiary information from Form 851 when available. The Joint Committee staff collects each valid parent-subsidiary pair for each year. In this step, parent-subsidiary relationships identified in each year are not transferred to future or prior years as subsidiaries may be bought, sold, or spun-off. Generally, the parent-subsidiary information contained on Form 851 is not exhaustive. In some cases, subsidiary EINs are not reported on any annual tax filings for a parent entity.

During this mechanical matching step, the Joint Committee staff also assigns an entity type to each parent-level EIN. The entity type for an EIN that has been linked to a parent entity return is simply the type of return. For example, if the parent entity filed a Form 1120 corporate tax return, the entity type would be "1120", and if the parent entity filed a Form 1065 partnership tax return, the entity type would be "1065." For EINs that have not been linked to an entity-level return, entity type can take on descriptions captured on payroll filings (*e.g.*, Form 941 or Form

[3] The parent entity IRS forms used in the creation of this bridge include Forms 1120, 1120-PC, 1120-REIT, 1120-RIC, 1120-S, 1120-C, 1120-F, 1120-FSC, 1120-H, 1120-ND, 1120-POL, 1120-L, 1065, 1065-B, 1040 Schedule C, 1040 Schedule F, 1066, 1041, 990, 990-C, 990-PF, 990-R, 990-T, and 990-ZR.

[4] Firms may have multiple EINs for a variety of reasons, including the possession of legacy EINs from mergers and acquisitions.

[5] The only restrictions for W-2s to be included as part of the matching process: the recipient and payer identification numbers must be valid, and the W-2 must contain at least $250 of Medicare wages (box 5).

W3) or internal IRS entity type classifications. Entity type classifications for these EINs can be as broad as "business" or as narrow as "religious." Subsidiary EINs are assigned the entity type of their parent.

For the final step of the mechanical matching process, the Joint Committee staff applies an interpolation procedure. This interpolation procedure fills in gaps for years in which no parent company documents a parent-subsidiary relationship on Form 851. The Joint Committee staff assumes that, for a given (non-parent) EIN in a given year, if (1) the EIN is not linked to a parent entity EIN and is not used to file a parent entity return, (2) the EIN was linked to a parent entity EIN in a previous year, and (3) the EIN is not listed on any other parent entity's Form 851, then the EIN is linked to the parent entity EIN that most recently listed the EIN.

The second step is a manual matching process. Manual matching targets the largest corporate entities (ranked by wages paid), with the goal of matching as many of the W-2 filings associated with these entities as possible. To begin, the Joint Committee staff calculates wage payments from the tax returns of these corporate entities. In the case of a corporation filing Form 1120,[6] wage payments are calculated by taking the sum of (1) a parent entity's cost of labor, found on line 3 of Form 1125-A, *Cost of Goods Sold*; (2) a parent entity's officer compensation, found on line 12 of Form 1120; and (3) a parent entity's salaries and wages, found on line 13 of Form 1120. The Joint Committee staff then collects the Form W-2 filings for each of a given parent entity's subsidiaries. From these forms, the Joint Committee staff takes the sum of Medicare wages (box 5) for all the Form W-2 filings associated with each parent entity.[7] In each year, the Joint Committee staff then examines all parent entities where less than 90 percent of their wage payments are accounted for by the sum of Medicare wages from matched Form W-2 filings and that have more than $1 billion in unmatched wage payments (a little over 100 firms per year). The Joint Committee staff then searches the population of unmatched (to other entity returns) W-2 filings (aggregated by EINs on Form W-2) for similar employer names, and manually amends the parent linkage.

The third step is an automated string-based matching process. This automated matching process is a fuzzy name matching procedure which compares similar but not identical parent entity names. In general, the fuzzy matching algorithm takes a given reported parent entity name and compares it to the employer names reported on each of the Form W-2 filings not matched using the mechanical or manual processes. See the Appendix for a detailed description of this approach.

[6] Analogous lines are retrieved for partnerships filing Form 1065.

[7] Box 5 wages are total Medicare wages and tips, the most complete measure of total compensation available on Form W-2 and the most comparable with items corporations might deduct as wages.

## II. EVALUATING THE BRIDGE

One assessment of the efficacy of the bridge involves matching the population of Form W-2s from 2018 and the corporate cross-section produced by SOI. The corporate cross-section is a large sample of edited corporate income tax returns. For the purposes of the bridge, any entity that files an entity-level tax return (e.g., Form 1120) is considered a parent entity, and as a result each observation in the SOI corporate cross-section is classified as a parent entity. The sampling weights within this data set allows the SOI corporate cross-section to be representative of the population of corporate entities.

To assess the value of the bridge for providing a complete picture of a firm's workforce, the Joint Committee staff matched the population of Form W-2 filings, both with and without the application of the bridge, to subchapter C corporations in the SOI corporate cross-section with positive wage deductions. The Joint Committee staff calculates firm size by number of employees in two different ways: as the number of unique Form W-2 filings aggregated either by the payer EIN recorded on form W-2, ignoring sister EINs belonging to the same firm ("naïve" W-2 aggregation), or by the parent EIN collected through the bridge ("bridged" W-2 aggregation).

Table 1 illustrates the effect of the bridge when producing counts of firms (where firm is used interchangeably with parent entity), employees, and wage deductions for subchapter C corporations in the 2018 SOI corporate sample. There are two rows for each set of statistics, which separately display statistics when using the bridge to connect W-2 EINs ("bridged") and when treating each W-2 EIN as if it represents a unique firm ("naïve"). The columns are different firm sizes, measured by number of employees, and a "total" column.[8]

The first set of rows displays the weighted number of corporations (in thousands) in the SOI cross-section that are linked with at least one W-2 reporting at least $5,000 of wages.[9][10] Differences in firm counts from the naïve column to the bridged column can result from (1) parent entities that are not matched to any W-2s being matched to one or more W-2s using the bridge, or (2) parent-entities that were previously matched to at least one W-2 being linked with more W-2s, causing them to move from one firm-size grouping to another. The second set of rows displays the number of employees (in millions) in each firm-size grouping, separately for the naïve and bridged approach. The third set of rows displays total wages paid (in billions) as measured by summing Medicare wages on each W-2.

[8] Parent entity returns do not contain information on number of employees. Employee counts are obtained by aggregating linked W-2s.

[9] The SOI corporate cross section is a sample of corporate tax returns. The sample is representative of the population of corporate tax returns when weights (proportional to the inverse of sampling probability) are applied.

[10] Approximately 1.5 million corporate returns (weighted) underlie this table, 59 percent of which have positive wage deductions, totaling $3.4 trillion. This wage total represents 40 percent of all wages paid in the United States in 2018 (by all employers, including government employers), as measured by Medicare wages reported on Form W-2.

**Table 1: Effect of Bridge on Employee Counts and Total Wages by Employment Size of Corporations (2018)**

| | | Employee Counts for Parent Entities Matched to W-2s | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1-4 | 5-9 | 10-19 | 20-99 | 100-499 | 500-9,999 | 10,000+ | Total |
| Firm Counts (thousands) | Naïve | 367.8 | 145.7 | 101.4 | 92.4 | 16.5 | 2.8 | 0.2 | 726.8 |
| | Bridged | 370.7 | 148.7 | 104.3 | 97.0 | 21.1 | 6.3 | 0.7 | 748.9 |
| Employees (millions) | Naïve | 0.8 | 1.0 | 1.4 | 3.6 | 3.3 | 4.4 | 5.1 | 19.5 |
| | Bridged | 0.8 | 1.0 | 1.4 | 3.9 | 4.3 | 12.3 | 28.7 | 52.4 |
| W-2 Wages (billions) | Naïve | $36 | $42 | $64 | $198 | $202 | $361 | $343 | $1,246 |
| | Bridged | $36 | $42 | $65 | $211 | $265 | $870 | $1,732 | $3,221 |

Notes: The naïve rows in this table were created from the population of W-2 filings merged with the edited SOI corporate tax return data. The bridged rows incorporate information from additional sources and filings and seeks to link related EINs. All data are from tax year 2018. W-2 wages are censored below at $250 (per W-2), and employee counts are censored below at $5,000 in W-2 wages from a parent entity.

Table 1 shows that using the parent-subsidiary bridge shifts the distribution of employees and wages paid toward large-employer firms. Relative to a naïve approach of assuming each firm has a single EIN, there are more than twice as many firms with 500 or more employees, more than four than times as many employees at these large-employer firms, and more than three times as many wages paid at large-employers. This shift is accomplished by the bridge in two ways. First, firms that do not match any W-2s with a direct EIN linkage are successfully linked to W-2s. This is most clearly seen in the "Total" column for firm counts, which increased from 726,800 firms in the naïve approach to 748,900 firms with the bridge. Second, direct EIN merges might only link a subset of employees, while the bridge more completely links employees back to the parent entity. As a result, some firms that appear to have few employees using the naïve approach are now identified to be a part of a larger firm using the bridge.

While the bridge improves linkages between entity and individual wage filings, it does not result in a perfect match. Timing differences between W-2 filings and entity tax return filings (*e.g.*, non-December fiscal year firms and those reporting cost of labor as part of the cost of goods sold) and different approaches to populating entity-level tax returns (*e.g.*, companies may report wage deductions on non-wage lines or include contractor and other non-employee compensation in wage lines) introduces a degree of disagreement between these two data sources, and the bridge continues to systematically under-count some amount of wages and employees. The under-counting associated with the bridge, however, is significantly less than the undercounting using the naïve approach. For 2018 data, the bridged approach matches wages totaling approximately 95 percent of wage compensation deducted on corporate tax returns whereas the naïve approach matches only roughly one-third of all wages deducted.

Table 2 provides a second set of insights into the validity of the bridge, by comparing the bridged W-2s with an external dataset and expanding from subchapter C corporations in the SOI cross-section to all non-governmental employers in the United States. This table compares firm counts (thousands), employee counts (millions), and wages paid (billions) as measured using the bridged W-2s and the Statistics of U.S. Businesses ("SUSB") data gathered and reported by the U.S. Census Bureau. The SUSB information is collected by the U.S. Census Bureau through a survey of more than six million individual establishments, and the SUSB works to connect establishments across multi-unit enterprises which are broadly analogous to the parent entities which the Joint Committee staff uses.[11] If the parent-subsidiary bridge correctly assigns employee Form W-2 filings to parent-entities, then the number of firms, number of employees, and amount of Form W-2 wages in each firm size bin should be similar to statistics reported by the SUSB. However, the data collection methods between the SUSB and tax filings are different – one is a survey of businesses conducted by the Census Bureau and the other is generally required by the IRS for purposes of administering the Internal Revenue Code – and some degree of difference is expected.

This table is structured in an analogous manner to Table 1. Each set of rows displays two sets of information: the counts or totals from the bridged W-2 data or the Census SUSB data. The first two rows of Table 2 report the number of firms within each firm size bin. The third and fourth rows compare employee counts by firm size, and the fifth and sixth rows compare payroll by firm size.

[11] For the SUSB, an establishment represents a discrete geographic entity where a single business operates.

**Table 2: Bridged Distribution of Firms, Employees, and Wages Compared to the Census Statistics of U.S. Businesses Survey (2018)**

| | | Firm Employee Counts | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 0-4 | 5-9 | 10-19 | 20-99 | 100-499 | 500-9,999 | 10,000+ | Total |
| Firm Counts (thousands) | Tax Data | 3,920 | 1,044 | 662 | 569 | 109 | 23 | 1 | 6,329 |
| | Census | 3,752 | 1,013 | 646 | 551 | 93 | 19 | 1 | 6,076 |
| Employees (millions) | Tax Data | 6.7 | 6.9 | 8.9 | 22.5 | 21.7 | 37.2 | 40.2 | 144.2 |
| | Census | 6.0 | 6.7 | 8.7 | 21.6 | 18.3 | 31.0 | 38.7 | 130.9 |
| Payroll (billions) | Tax Data | $244 | $246 | $337 | $936 | $1,046 | $2,091 | $2,351 | $7,252 |
| | Census | $287 | $261 | $351 | $961 | $960 | $1,891 | $2,388 | $7,097 |

Notes: The bridged rows are constructed from W-2s and tax infomation, the Census rows are retrieved from the SUSB survey. In the tax data, payroll is total W-2 wages, which are censored below at $250 (per W-2), and employee counts are censored below at $5,000.

The statistics on firm counts, employees, and wages calculated from the aggregation of Form W-2 filings closely resemble those derived from the SUSB.[12] The tax data contain more firms, employees, and wages. These differences are likely due to the SUSB capturing a point in time, while tax data are annual measures, and that certain entities are excluded from the survey that are included in the tax data.[13] But in general, this comparison provides external validity to the data and methods underlying the construction of the bridge.

[12] U.S. Census Bureau (2020). Statistics of US Businesses (SUSB). Retrieved from: https://www.census.gov/data/tables/2018/econ/susb/2018-susb-annual.html.

[13] In particular, the SUSB excludes entities engaged in crop and animal production, rail transportation, pension funds, trusts, estates, office of notaries, private households, and public administration.

One issue affecting the tax data relative to the SUSB is the existence of Certified Professional Employer Organizations ("CPEOs"), which are organizations that handle various payroll tax reporting responsibilities on behalf of their client businesses. The IRS allows these organizations to file W-2s on behalf of their clients using the CPEO's EIN rather than the employer's EIN. Because CPEOs often use the same EIN to file W-2s for multiple firms, it is not possible to determine which company employs someone receiving a W-2 from a CPEO. In the corporate cross section, only 85 percent of firms with less than $10 million in wages are linked to a W-2 and, to the extent clients of CPEOs are primarily smaller firms, this may be evidence of their effect.

## III. DISCUSSION OF IMPLICATIONS FOR JOINT COMMITTEE REVENUE ESTIMATES

The Joint Committee staff has long relied on detailed tax return data to provide Congress with analyses of the economic and budgetary impact of proposed legislation. The creation of the parent-subsidiary bridge described in this document substantially improves the precision of such analyses, as they relate to the characteristics of firms' workforces. This section outlines several examples of revenue estimates which have benefited from the bridge, including employee retention credits and restrictions on compensation deductions.

In 2020, Congress enacted several policies in which eligibility was linked to the number of employees working for a firm. One such example was section 2301 of the CARES Act,[14] which created the Employee Retention Credit for Employers Subject to Closure due to COVID-19. This provision allowed eligible employers to claim a credit for 50 percent of qualified wages in each calendar quarter against the employer share of Old-Age, Survivors, and Disability Insurance ("OASDI") taxes. The determination of qualified wages depended critically on the number of full-time and full-time-equivalent employees that the employer had in calendar year 2019. For eligible employers with fewer than 100 employees, qualified wages with respect to the credit included the entirety of wages paid by the employer in the given quarter. Conversely, employers with 100 or more employees could only claim wages paid to employees with respect to which the employee was not providing services. While W-2 filings do not allow for the direct measurement of full-time or full-time equivalent employees, the data are useful in determining the scope of the provision. Table 1 shows that reliance on naïve W-2 counts would understate the number of entities with less than 100 employees by 1.1 percent relative to the bridged W-2 counts. This would lead to an undercount of employees and wages by 4.0 percent and 2.9 percent, respectively.

In March 2021, Congress enacted the American Rescue Plan Act of 2021 ("ARPA").[15] One provision of this legislation modifies section 162(m) of the Internal Revenue Code (the "Code"), which prohibits publicly traded companies from deducting compensation in excess of $1 million for certain covered employees. Prior to ARPA, the covered employees were generally the chief executive officer ("CEO"), the chief financial officer ("CFO"), and the three other highest-compensated officers of the company. The provision expands the definition of "covered employee" to include the next five highest-compensated employees of the company, resulting in at least ten covered employees for each tax year. The parent-subsidiary bridge, which is particularly useful for linking Form W-2 compensation attributable to large entities with complicated group structures (*e.g.*, publicly traded companies) to associated entity-level tax filings, improves the Joint Committee staff's ability to estimate the amount of compensation deduction disallowed under the provision.

Congress has at times enacted employee retention credits against employer income tax in response to natural disasters. These provisions generally provide a credit equal to 40 percent of the wages (up to a maximum of $6,000 in wages per employee) paid by certain employers

[14] Pub. L. No. 116-136.

[15] Pub. L. No. 117-2.

harmed by the applicable disaster to employees employed in the applicable disaster zone during the period when the employer’s business was inoperable due to the applicable disaster.[16] In these cases, observing the location of the entity filing a tax-return is not sufficient to determine whether they may have eligible wages, particularly in the case of multi-establishment firms which may have some employees inside disaster zones and some outside. The Form W-2 filings alone allow observation of the location of a given employee but give little insight into the tax status of their employer and how likely the employers are to be able to make use of such credits. The linkage between the two allows for a more complete analysis of these credits, and the bridge is especially useful since it improves the linkage primarily for larger, more complex firms who are likely to operate in more than one establishment in more than one geographic area.

[16] One such example is Code section 1400R(a) which was enacted to provide relief for employers affected by Hurricane Katrina.

## APPENDIX: DESCRIPTION OF FUZZY NAME-MATCHING PROCEDURE

This appendix describes the fuzzy name-matching procedure the Joint Committee staff uses to augment the linkages in the parent-subsidiary bridge. The data used are comprised of entity-level tax returns (*e.g.*, Forms 1120, 1120-S, 1065) and W-2 EINs that are not already linked to a parent EIN and are not associated with a government, religious institution, or non-profit. Entity level returns and W-2s aggregated by EINs with less than two million dollars in wage payments are excluded from the procedure. The Joint Committee staff runs two similar name matching procedures to identify possible matches.[17] The first splits strings into “bigrams” which are consecutive pairs of characters such that a string such as “John Doe” is split into “Jo”, “oh”, “hn”, “n_” “_D”, “Do”, “oe”. Each payer name reported on a W-2 is paired with each entity name reported on an entity-level tax return. Each pair of these names is given a score based on the percentage of bigrams that match between the two, with bigrams inversely weighted by the frequency with which they appear in the combined datasets:

$$weight\ of\ bigram\ b = \frac{1}{n_b}$$

where $n_b$ is the number of times bigram *b* appears in the data.

The score used for the bigram procedure is called a Jaccard score and takes the form:

$$Jaccard\ score = \frac{m}{\sqrt{s_1 * s_2}}$$

where *m* is the weighted count of matching bigrams and $s_1$ and $s_2$ are the weighted counts of the bigrams in the W-2 name ($s_1$) and the entity name ($s_2$).

For example, if one were comparing “John Doe” with “Jane Doe” (“Ja”, “an”, “ne”, “e_”, “_D”, “Do”, “oe”), since the bigrams “Jo”, “oh”, “hn”, “n_”, “Ja”, “an”, “ne”, and “e_” appear only once, and the bigrams “D_”, “Do”, and “oe” appear twice, the Jaccard score would be:

$$Jaccard\ score = \frac{\overset{\text{"\_D"}}{\frac{1}{2} * 1} + \overset{"Do"}{\frac{1}{2} * 1} + \overset{"oe"}{\frac{1}{2} * 1}}{\sqrt{\left(\frac{"Jo"}{1} + \frac{"oh"}{1} + \frac{"hn"}{1} + \frac{\text{"n\_"}}{1} + \overset{\text{"\_D"}}{\frac{1}{2} * 1} + \overset{"Do"}{\frac{1}{2} * 1} + \overset{"oe"}{\frac{1}{2} * 1}\right) * \left(\frac{"Ja"}{1} + \frac{"an"}{1} + \frac{"ne"}{1} + \frac{\text{"e\_"}}{1} + \overset{\text{"\_D"}}{\frac{1}{2} * 1} + \overset{"Do"}{\frac{1}{2} * 1} + \overset{"oe"}{\frac{1}{2} * 1}\right)}} = \frac{\frac{3}{2}}{\sqrt{\frac{11}{2} * \frac{11}{2}}} = \frac{3}{11}$$

[17] Prior to performing the fuzzy matching procedure, the payer name field on Form W-2 and the name field on parent entity tax returns are cleaned using the same procedure: commonly observed words or acronyms are dropped or standardized, punctuation is removed, and trailing or leading spaces are removed.

The Joint Committee staff also “ignores” bigrams that appear with a frequency greater than 0.2 instances per observation. Applying this rule to the example above would mean that no match is produced since there are only two observations, so each bigram has a frequency of at least 0.5 instances per observation.

The second procedure splits strings into groups of characters separated by a space called “tokens” instead of bigrams. For the “John Doe” example, one would get “John” and “Doe”. The weights are calculated in the same way, but the Joint Committee staff uses a “simple” score instead of a Jaccard score that is calculated as:

$$Simple\ score = \frac{2 * m}{t_1 + t_2}$$

where $m$ is the number of tokens that match between the two words, $t_1$ is the weighted count of tokens in the W-2 payer name and $t_2$ is the weighted count of tokens in the entity name.

For the “John Doe” and “Jane Doe” example that would be:

$$Simple\ score = \frac{2 * \overset{\text{"Doe"}}{\frac{1}{2}}}{\left(\overset{\text{"John"}}{1} + \overset{\text{"Doe"}}{\frac{1}{2}}\right) + \left(\overset{\text{"Jane"}}{1} + \overset{\text{"Doe"}}{\frac{1}{2}}\right)} = \frac{1}{3}$$

Jaccard scores and simple scores differ on how much they penalize “extraneous” units. An “extraneous” unit is a unit that appears in one name, but not in the other. In the case where the weighted number of units is the same, the simple score and Jaccard score will be equivalent. For example, if one were to calculate a Jaccard score for the token procedure one would get:

$$Jaccard\ score = \frac{\overset{\text{"Doe"}}{\frac{1}{2}}}{\sqrt{\left(\overset{\text{"John"}}{1} + \overset{\text{"Doe"}}{\frac{1}{2}}\right) * \left(\overset{\text{"Jane"}}{1} + \overset{\text{"Doe"}}{\frac{1}{2}}\right)}} = \frac{\frac{1}{2}}{\sqrt{\frac{3}{2} * \frac{3}{2}}} = \frac{\frac{1}{2}}{\frac{3}{2}} = \frac{1}{3}$$

Adding a middle name “Jay” to “John Doe” but not to “Jane Doe” would produce the scores:

$$Simple\ score = \frac{2 * \overset{\text{"Doe"}}{\frac{1}{2}}}{\left(\overset{\text{"John"}}{1} + \overset{\text{"Jay"}}{1} + \overset{\text{"Doe"}}{\frac{1}{2}}\right) + \left(\overset{\text{"Jane"}}{1} + \overset{\text{"Doe"}}{\frac{1}{2}}\right)} = \frac{1}{\frac{5}{2} + \frac{3}{2}} = \frac{1}{4}$$

$$Jaccard = \frac{\overset{"Doe"}{\frac{1}{2}}}{\sqrt{\left(\overset{"John"}{1} + \overset{"Jay"}{1} + \overset{"Doe"}{\frac{1}{2}}\right) * \left(\overset{"Jane"}{1} + \overset{"Doe"}{\frac{1}{2}}\right)}} = \frac{\frac{1}{2}}{\frac{\sqrt{15}}{2}} = \frac{1}{\sqrt{15}} < \frac{1}{\sqrt{16}} = \frac{1}{4}$$

The choice of a Jaccard score for the bigram procedure and a simple score for the token procedure stems from a belief that an additional or mismatched token provides more evidence against a match than does an additional or mismatched bigram.

After each pair consisting of a W-2 payer name and an entity name is given a bigram Jaccard score and a token simple score, a simple average is taken between the two scores to create an "average" score. Pairs are then eliminated first if the compensation linked to the W-2 EIN is more than 1.1 times the amount of wages deducted on the entity-level return associated with the entity EIN. A pair is also dropped if the average score is below 0.45 or if the bigram Jaccard score is below 0.85. Finally, matches are limited to those with a bigram Jaccard score over 0.95, a token simple score over 0.98, an average score of 0.9, or an average score above 0.8 with W-2 compensation over four million dollars. In the event of a W-2 EIN being matched to multiple entity EINs, the match to the entity EIN that deducted the most wages is kept. Finally, matches with a token bigram score of zero are dropped if their bigram Jaccard score is lower than 0.995.