Using Entity Resolution to Improve Inward FDI—QCEW Estimates

Lowell G. Mason
Data Scientist
Employment Research and Program Development Staff, OEUS

FCSM 2023
October 26, 2023

Inward FDI-QCEW Estimates

- Inward FDI-QCEW Estimates:
  - Joint BLS and BEA project to identify QCEW establishments that were foreign-owned during 2012.
  - QCEW *establishment* data augment BEA *enterprise-level* data:
    - More granular geographic and industry estimates, and
    - Adding employment and wage estimates by occupation.
  - https://www.bls.gov/fdi/home.htm

2012 Inward FDI-QCEW Matching Process

- Initial link using common identifier: EIN

- Manual review by program analysts to:
  - Remove EINs that were linked in error.
  - Add EINs that were not linked but should have been.

Inward FDI-QCEW Matching Challenges

- EINs are a poor identifier:
  - Employer Tax ID Numbers (EINs) are neither consistent nor unique:
    - Many employers use multiple EINs (possibly for multiple purposes)
    - EINs are likely reported to BEA and BLS by different respondents
  - Initial matching error rate using EIN: 87.7%

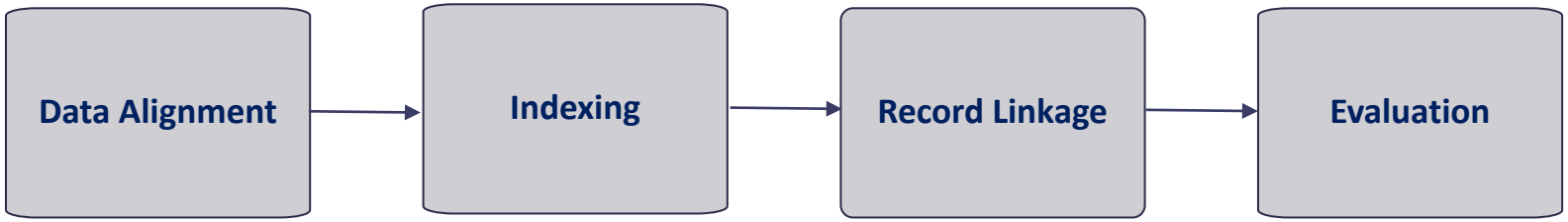
Inward FDI-QCEW Matching Challenges, Cont.

- Analyst review was effective but very labor intensive:
  - Final matching error rate reduced to 19.0%.
  - Labor costs exceeded 1,500 hours for the initial review by BLS program analysts.
  - Additional review by BEA and BLS subject matter experts (wasn't timed).

Why Use Entity Resolution?

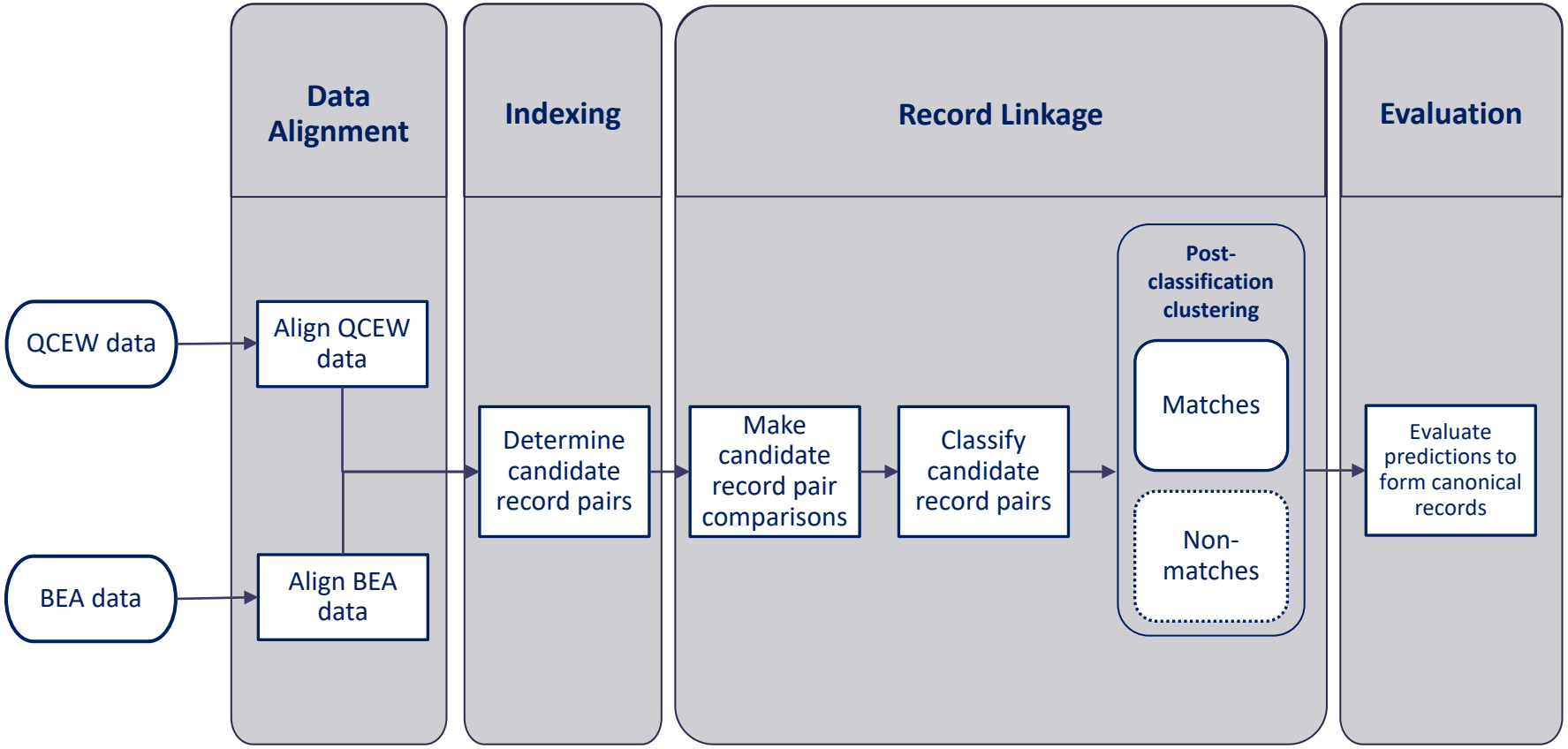
- Entity resolution is a collection of statistical and computational methods used to link related records across multiple data sources that lack consistent and unique IDs
  - Increases the effectiveness of the initial, automated steps
  - Better initial matches require less manual review

Entity Resolution Pipeline



- *Data alignment* is transforming records and/or their attributes so that they align across the data sources.
- *Indexing* maps similar records into partitions so only the records within one partition will be considered.
- *Record linkage* is the process of merging multiple data sources and removing duplicate records across data sources.
- *Evaluation* concludes the process by resolving inconsistencies that remain amongst linked records to create a single, canonical record.

Detailed entity resolution process

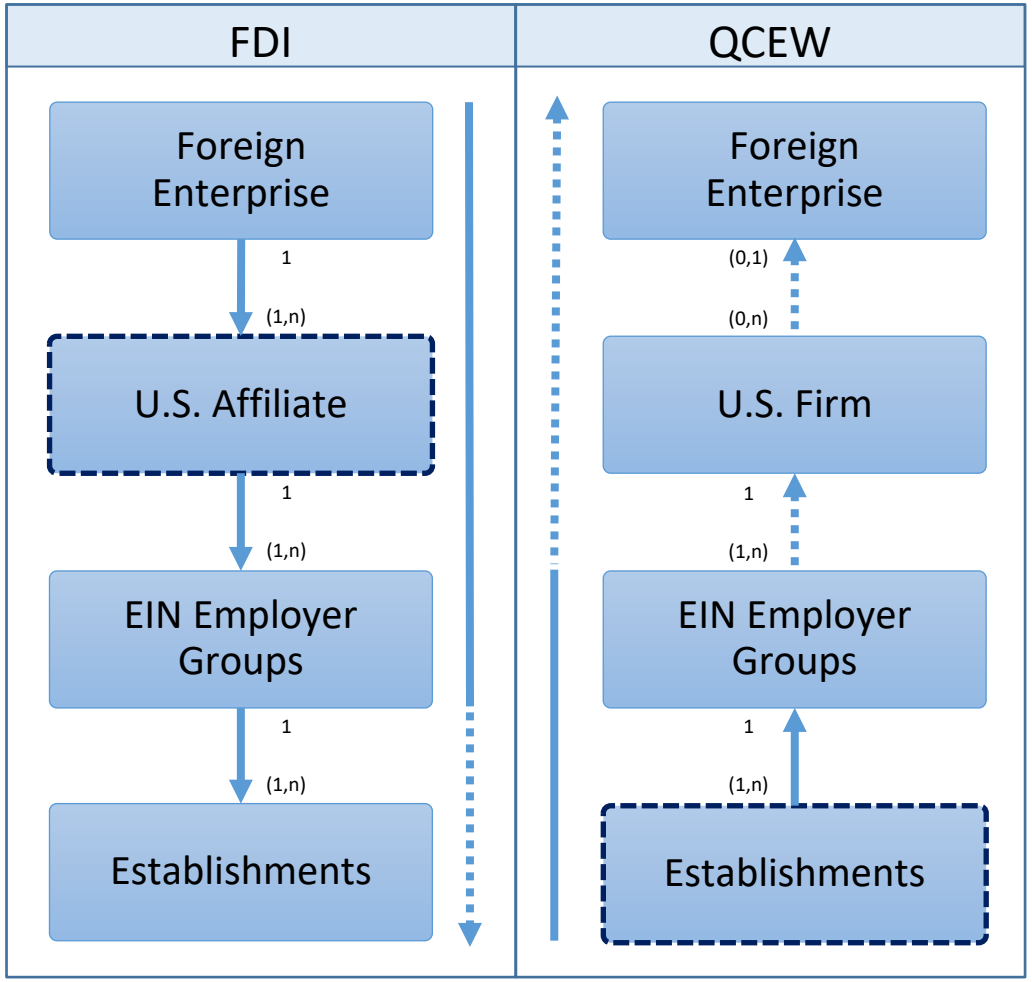


Data Alignment

- Entity resolution is complicated by the fact that an affiliate can link to more than one establishment (and more than one EIN)
  - To better align the data sources, we aggregate establishments to EIN employer groups.
    - Still one-to-many, but less so
  - Calculate summary measures at the EIN level:
    - Number of states and establishments
    - Establishment employment and total wages
    - Distribution of establishments by states and industries
    - Lists of names and addresses

Employer Structure Inconsistencies

- The data sources are not consistent.
- This is due to differences in the unit of measures:
  - FDI: affiliate
  - QCEW: establishment



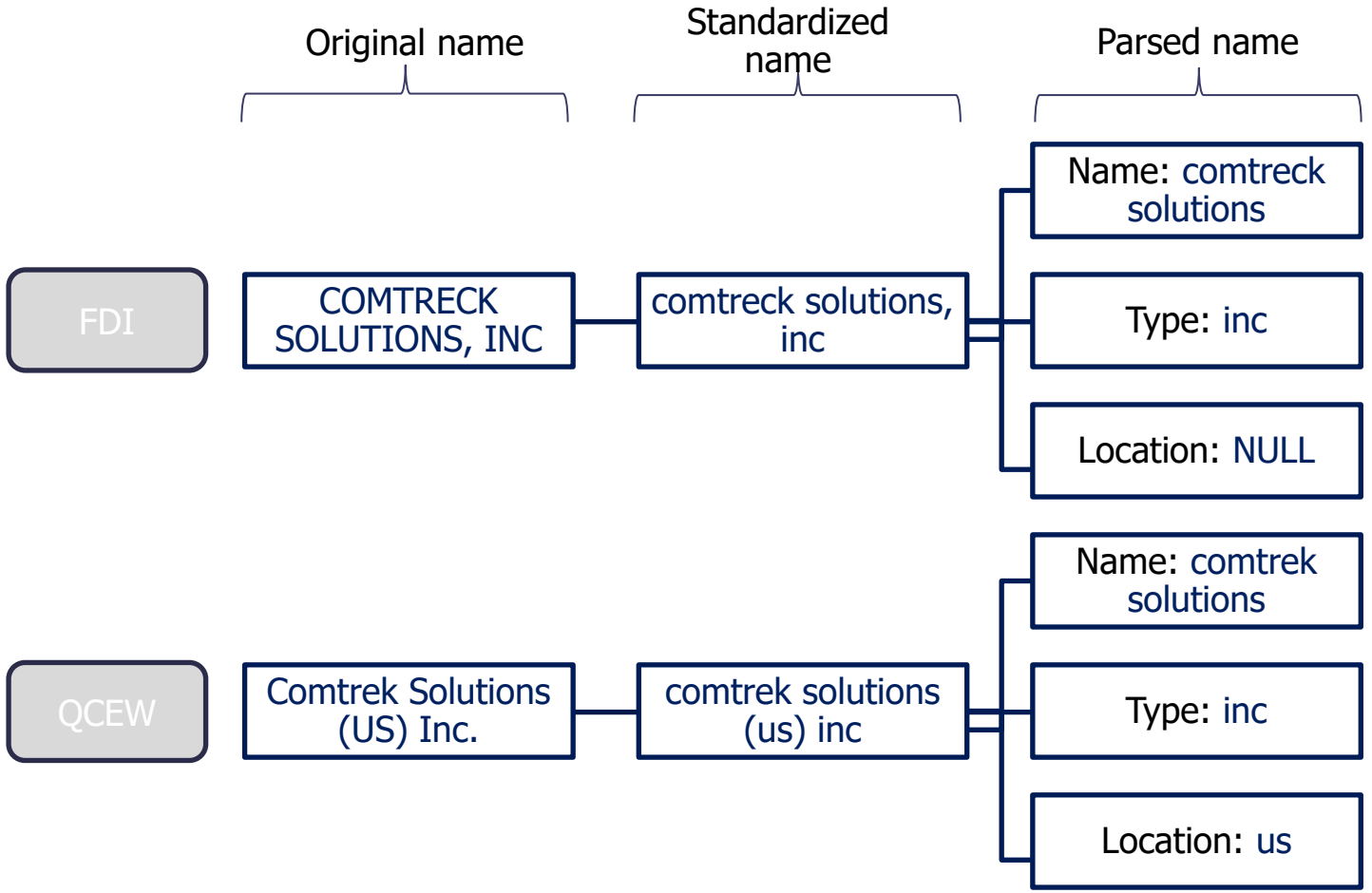
Features Common to Both Data Sources

FDI Affiliate		QCEW Establishment	
Feature details	**Dimension**	**Features details**	**Dimension**
Primary and secondary EINs for the affiliate as well as an EIN for each subsidiary (e_a=0…m)	(2 + e_a, 1)	Establishment EIN	1
Affiliate employment, total and broken out by state	(52, 1)	Establishment employment	1
Affiliate total compensation	1	Establishment total wages	1
4-digit NAICS, total and broken out by the (10, 4, or 0) largest industries by sales (depending on affiliate size)	1	6-digit NAICS	1
Affiliate name and subsidiary names (n_a=0…m)	(1+n_a, 1)	Trade and legal establishment names (n_e=1 or 2)	(n_e, 1)
Headquarters and/or mailing addresses (a_a=1…2) by component (street, city, state, zip)	(a_a, 4)	Physical, mailing, and/or other addresses (a_e=1…3) by component (street, city, state, zip)	(a_e, 4)
Contact information (Name, phone, fax, email)	(1, 4)	Contact information (Name, phone, fax, email)	(1, 4)
Affiliate type (Corporation, LLC, Partnership, Individual, Other)	1	Establishment type (Corporation, Partnership, Individual, Other)	1

Aligned Features Common to Both Data Sources

FDI Affiliate		QCEW EIN Employer Group ($i = 1, \ldots, N$ establishments in group)	
Feature details	**Dimension**	**Features details**	**Dimension**
Primary and secondary EINs for the affiliate as well as an EIN for each subsidiary (e_a=0…m)	(2 + e_a, 1)	All establishment EINs	($\sum_{i \in N} e_{ei}$, 1)
Affiliate employment, by state	(52, 1)	Aggregate EIN employment, by state	(52, 1)
Affiliate total compensation	1	Establishment total wages	1
1 if NAICS sector (or subsector), 0 otherwise	(20, 1)	Aggregate EIN employment shares, by NAICS sector (or subsector)	(20, 1)
Affiliate and subsidiaries names (n_a= 0,…,N_a)	(n_a+1, 1)	All establishment trade and legal names (n_{ei}= 1 or 2)	($\sum_{i \in N} n_{ei}$, 1)
Headquarters and/or mailing addresses (a_a=1…2) by component (street, city, state, zip)	(a_a, 4)	All establishment physical, mailing, and/or other addresses (a_{ei}= 1,…,3) by component (street, city, state, zip)	($\sum_{i \in N} a_{ei}$, 4)
Contact information (name, phone, fax, email)	(1, 4)	Contact information (name, phone, fax, email)	(*N*, 4)
1 if affiliate is of type t_a, (t_a = {corporation, partnership, individual, other}), 0 otherwise	(1, 4)	Aggregate EIN establishment type shares	(*N*, 4)

Data Alignment Example



Indexing

- The record linkage search space is all the possible combinations of records in the two data sources:
  - $|FDI \times QCEW| = mn \approx 6.8 \times 10\text{^}10$
  - The indexing step aims to reduce this search space so that it is computationally tractable. The result is the set of all *candidate pairs*.
- There are two methods:
  - Deterministic (also referred to as blocking)
  - Probabilistic (e.g., similarity search, Approximate Nearest Neighbors, etc.)

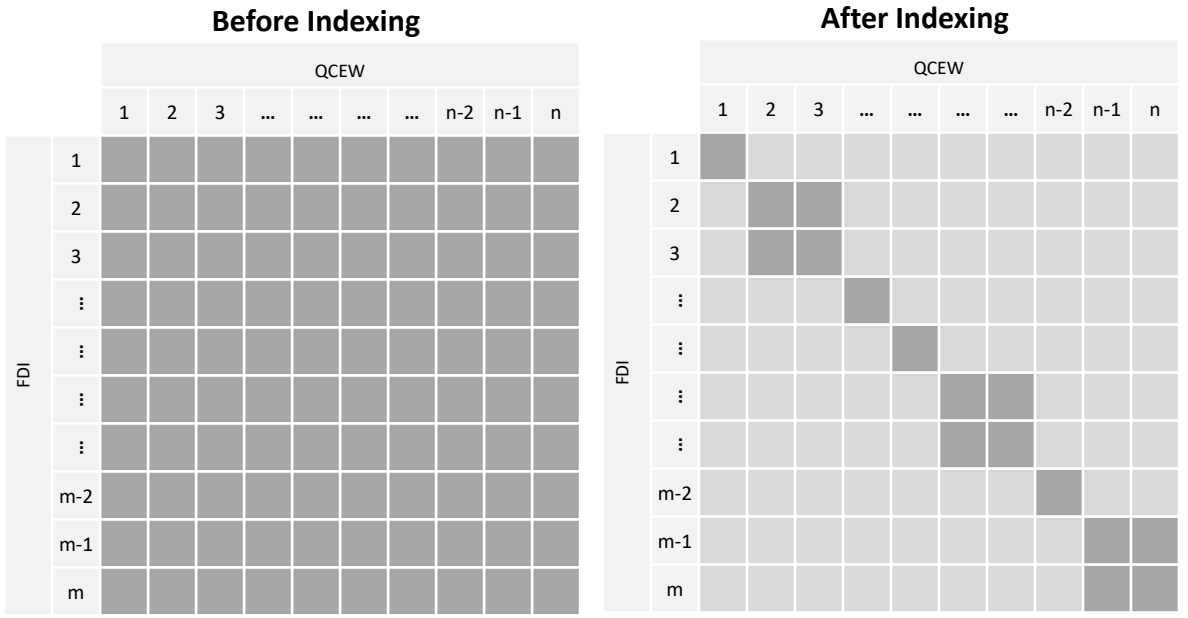
Deterministic Indexing

- Partition the search space into blocks by requiring some subset of the attributes for records in each data source to match according to some function.

Data sources:
- FDI: $(i = 1, \dots, m$ observations$)$
- QCEW: $(j = 1, \dots, n$ observations$)$

Size of search space:
- Before indexing: mn
- After indexing: 16% of mn*

*Assuming a 1-to-1 correspondence between area and the number of observations



Indexing, Cont.

- Hybrid approach:
  - Block by state
  - Use Locality Sensitive Hashing within blocks
- Evaluation:
  - Search space reduced from $\sim 6.8 \times 10^{10}$ to 2,172,330 candidate pairs, or 0.003195% the size of the original.
  - Of the well-matched pairs, 90.90% are in the reduced search space.
  - Class imbalance ratio: $\sim \frac{1}{115}$.

Record Linkage

- The record linkage step takes the candidate pairs returned from indexing and determines if the candidate pairs should or should not be linked
- Record linkage sub-processes:
  - Candidate pair comparison: measure how “similar” each candidate pair is across the common attributes
  - Classification: predict if candidate pair represents a match or non-match using the similarity measures
  - Post-processing clustering: use the results of the classification methods to impose linkage constraints to ensure a coherent output

Candidate Pair Comparison: Similarity Measures

- Normalized similarity measures:

  Given a common attribute, $a_i, i \in 1 \ldots n$, for records u and v, $u \in 1 \ldots U$ and $v \in 1 \ldots V$, the normalized similarity function s_{iuv} is:

  $$s_{iuv} = sim(a_{iu}, a_{iv}) \rightarrow [0, 1].$$

  Larger values of s_{iuv}denote greater similarity.

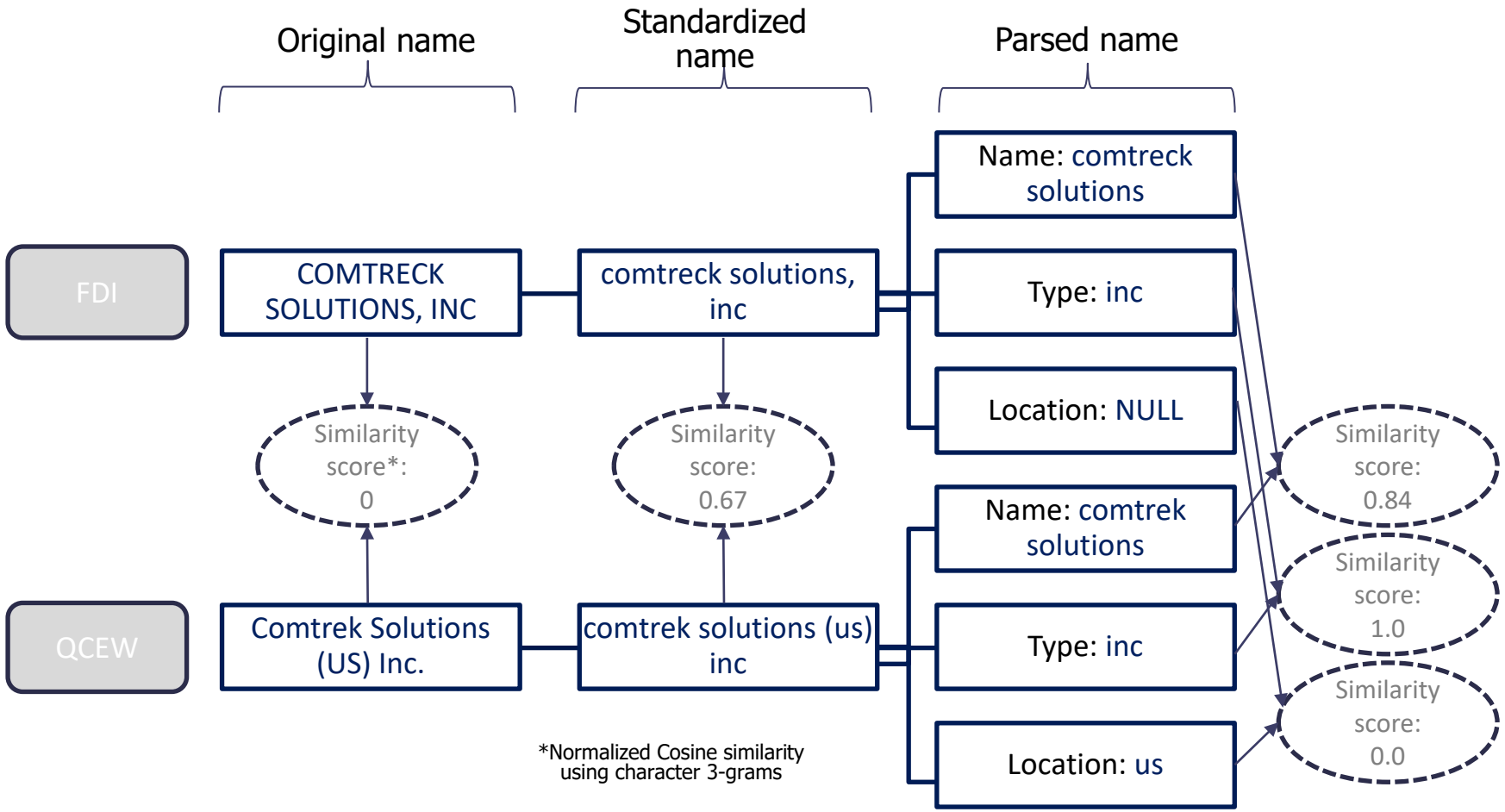
- For a normalized distance metric, d, the corresponding similarity measure is $\mathrm{s} = 1 - \mathrm{d}$.

- The vector of similarity measures for all common attributes is referred to as the comparison vector.

Types of Similarity Measures

- Numeric attributes: are generally calculated from a normalized distance measure, such as Manhattan ($l1$) or Euclidean ($l1$) distance.
- Categorical attributes: are compared using set-based similarity measures, such as Jaccard Similarity.
- String attributes:
  - Using common distance measures (e.g., Levenshtein, Hamming)
  - Sparse (e.g., TF-IDF) or dense (e.g., embedding methods) vector representation + Cosine similarity.

Candidate Pair Comparison Example



Similarity Measure Alignment

- For establishment attributes for which summary measures or distributions are calculated, similarity measures can be used directly.
- However, for attributes for which lists are created, aggregate similarity measures are required:
  - For example, both affiliates and EIN aggregations may have multiple employer names:
    - $s_{|name|,uv} = \left| sim\left(a_{name_i,u}, a_{name_j,v}\right)\right|_{\forall\, name_i \in u,\, name_j \in v}$

Applicability of Similarity Measures

FDI Affiliate		QCEW EIN Employer Group ($i = 1, \ldots, N$ establishments in group)		Similarity Measures Are Applicable?		Number of Similarity Measures
Feature details	Dimension	Features details	Dimension	Directly	Using an aggregation?	
Affiliate employment, by state	(52, 1)	Aggregate EIN employment, by state	(51, 1)	Yes		1
1 if NAICS sector (or subsector), 0 otherwise	(20, 1)	Aggregate EIN employment shares, by 3-digit NAICS	(20, 1)	Yes		1
Affiliate total compensation	1	Aggregate EIN total wages	1	Yes		1
Affiliate and subsidiaries names (n_a= 0,…,N_a)	(n_a+1, 1)	All establishment trade and legal names (n_{ei}= 1 or 2)	$(\sum_{i \in N} n_{ei}, 1)$	No	Yes	1
Physical address, by component (street, city, state, zip)	(1, 4)	All establishment physical, mailing, and/or other addresses (a_{ei}= 1,…,3), by component (street, city, state, zip)	$(\sum_{i \in N} a_{ei}, 4)$	No	Yes	4
Contact information (name, phone, fax, email)	(1, 4)	Contact information (name, phone, fax, email)	(N, 4)	No	Yes	4
1 if affiliate is of type t_a, (t_a = {corporation, partnership, individual, other}), 0 otherwise	(1, 4)	Aggregate EIN establishment type shares	(N, 4)	No	Yes	4

Classification

- Classification aims to predict the linkage type for the candidate pairs, where linkage type can contain 2 or 3 class labels:
  - Match
  - Non-match
  - Potential match (optional)
- Several record linkage classification methods are used in practice, including:
  - Deterministic, rule-based methods,
  - Probabilistic record linkage methods and its modern extensions, and
  - Cluster-based methods.

Supervised Classification

- Supervised learning classification methods such as logistic regression, decision trees, random forests are commonly employed when labeled data are available:
  - Class label imbalance must be carefully considered:
    - Adequate amounts of training data, carefully selected training data, or modifications to the supervised methods that account for imbalance are often required.
  - One benefit to supervised methods is features that are pertinent to the individual data sources can be accounted for in addition to the comparison vector during modeling.

Supervised Classification, Cont.

- Supervised learning to classify candidate pairs as either a match or non-match.
  - Training data is the set of 2012 affiliates that are considered well-matched.
  - Modeled using Balanced Random Forests.
  - Precision used to optimize model hyper-parameters.
  - The trained 2012 model is then applied to 2017 data.

Supervised Classification Results

- Classification:
  - LF:
    - Balanced accuracy: 0.96
    - Recall: 0.94
    - Precision: 0.39
  - SF:
    - Balanced accuracy: 0.99
    - Recall: 0.94
    - Precision: 0.27
  - Mini:
    - Balanced accuracy: 0.98
    - Recall: 0.97
    - Precision: 0.53
  - CF:
    - Balanced accuracy: 0.94
    - Recall: 0.94
    - Precision: 0.12

Confusion Matrix		Actual Linkage	
		Not a match	Match
Predicted Linkage	Not a match	LF: 97,970 (0.98%) SF: 62,382 (0.99%) Mini: 62,410 (1.00%) CF: 45,483 (0.94%)	LF: 1,938 (0.02%) SF: 895 (0.01%) Mini: 301 (0.00%) CF: 2,667 (0.06%)
	Match	LF: 79 (0.06%) SF: 23 (0.06%) Mini: 10 (0.03%) CF: 24 (0.06%)	LF: 1,239 (0.94%) SF: 338 (0.94%) Mini: 345 (0.97%) CF: 348 (0.94%)

Post-Classification Clustering

- Aims to impose linkage constraints to ensure a coherent output.
  - Use the results of the classification methods (e.g., pairwise similarity measures, match probabilities, or the estimated match/non-match class labels).
- Various clustering methods have been proposed, including correlation clustering and hierarchical agglomerative clustering.

Post-Classification Clustering, Cont.

- Classification considers inward FDI affiliate-EIN candidate pairs individually.
  - By targeting precision during classification, there should be few missed matches, but many of the predicted EIN matches are false positives.
- We further refine the set of predicted matched EINs for an affiliate so that:
  - The refined set is composed of EINs with high match probabilities and with large pairwise similarity measures, and
  - The aggregate employment of the refined set is close to affiliate employment in total and distributed by state.

Post-Classification Clustering Results

Form type	Number of affiliates	Well-matched affiliates		Match error rate	
		Post-classification clustering?		Post-classification clustering?	
		No	Yes	No	Yes
LF	311	294 (94.5%)	311 (100.0%)	9.97%	4.49%
SF	125	118 (94.4%)	125 (100.0%)	11.54%	5.55%
Mini	63	40 (63.5%)	63 (100.0%)	219.88%	7.49%
CF	147	61 (41.5%)	147 (100.0%)	69.47%	6.54%

Evaluation

- All records from the data sources that refer to the same entity as identified in the record linkage step are merged to form a single representative record
- These may then be used for downstream tasks
- Also known as canonicalization

Ongoing Entity Resolution Research

- ASA/NSF/BLS Senior Research Fellowship:
  - "Improvements in Implementations and Analysis of Record Linkage Algorithms," by Roee Gutman (Brown University) funded in 2022.
  - Focused on improving matching BEA and BLS data.
- Reformulates entity resolution from a Bayesian perspective:
  - Classify EINs collectively.
  - Quantify the uncertainty associated with the linked data.

Ongoing Entity Resolution Research, Cont.

- Analyst review involves:
  - Evaluating (potentially) a lot of data
  - Some system interaction to correct invalid matches and add missed matches.
- Developing a web-based app for efficient review

Contact Information

Lowell G. Mason
Data Scientist
OEUS/ERPDS
www.bls.gov/ers
202-691-6244
mason.lowell@bls.gov