A Deep Dive into Entity Resolution

Rebecca C. Steorts

Department of Statistical Science, affiliated faculty in
Computer Science, Biostatistics and Bioinformatics, the
information initiative at Duke (iiD) and
the Social Science Research Institute (SSRI)
Duke University and U.S. Census Bureau

This work is supported by NSF CAREER Award 1652431 and
the Alfred Sloan Foundation (DRB #: CBDRB-FY20-309).

June 27, 2021

Entity resolution (record linkage or de-duplication) joins multiple data sets removes duplicate entities often in the absence of a unique identifier.

Entity Resolution

Why is entity resolution difficult?

Goals of Entity Resolution

Suppose that we have a total of M records in D data sets.

1. We seek models that are much less than $O(M^2)$ (quadratic).
2. We seek models that are reliable, accurate, fit the data well, and account for the uncertainty of the model.

Goals of Entity Resolution

Suppose that we have a total of M records in D data sets.

1. We seek models that are much less than $O(M^2)$ (quadratic).
2. We seek models that are reliable, accurate, fit the data well, and account for the uncertainty of the model.

These two goals fundamentally go against one another, making record linkage a very challenging problem.

Goals of Entity Resolution

Suppose that we have a total of M records in D data sets.

1. We seek models that are much less than $O(M^2)$ (quadratic).
2. We seek models that are reliable, accurate, fit the data well, and account for the uncertainty of the model.

These two goals fundamentally go against one another, making record linkage a very challenging problem.

Terminology

1. De-duplication
2. Record linkage
3. Entity resolution
4. Blocking

De-duplication

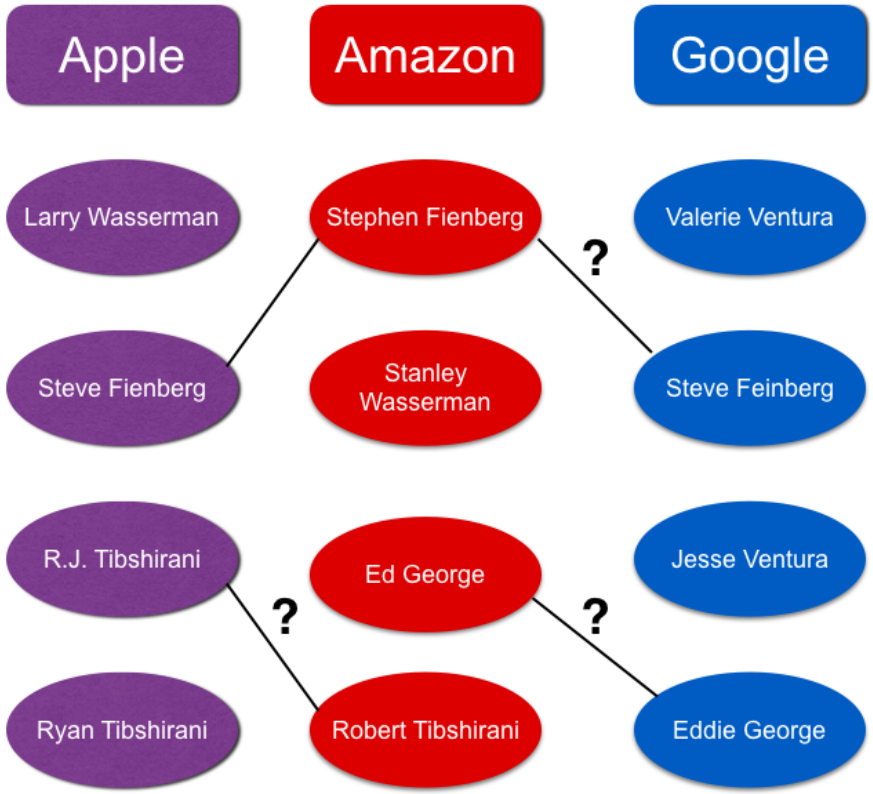
De-duplication

Much of the literature can be grouped into the case of de-duplication.

Common examples from both academia and industry are the following:

logistic regression, random forests, support vector machines, Bayesian adaptive regression trees, and locality sensitive hashing.

The entity resolution graph



De-duplication refers to duplication within just one database.
Record linkage refers to duplication across multiple databases.
Entity resolution refers to both duplication across and within databases (the entire graph).

Blocking

Often one performs blocking due to the fact that record linkage problems require a quadratic number of comparisons.

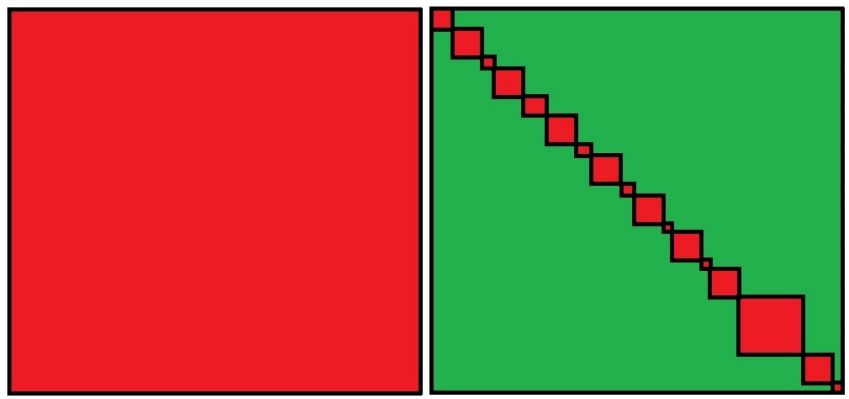


Figure 1: All-to-all record comparisons (left) versus partitioning records into blocks and comparing records only within each partition (right).

We will assume some method of blocking is embedded within a record linkage procedure.

Blocking

The most common method used for blocking is typically

1. deterministic blocking method
2. probabilistic blocking method

Examples include blocking on features (deterministic) or probabilistic types such as locality sensitive hashing.

See Christen (2012); Steorts, Ventura, Sadinle, Fienberg (2014); Chen, Shrivastava, Steorts (2018).

Common Methods for Entity Resolution

- Match on a unique identifier if it exists.
- Perform exact matching.
- Perform a likelihood ratio or hypothesis test.

[Newcombe (1959), Fellegi and Sunter (1969)].

Unique Identifier

Suppose that each feature has a unique identifier that we are sure is accurate, like social security number.

Then we can unique match records based on the unique identifier.

Problems occur this unique identifier is missing or has noise in it, etc.

Exact Matching

In exact matching, we compare all features. We decide if the record is a match if they agree on all features. Otherwise, we decide the record is a non-match.

240 Collins Drive
Pittsburgh, PA
50-54
412-793-3313

537 N Neville Street
Pittsburgh, PA 15213
65+
412-683-5599

Why would this method be bad for evaluation purposes?

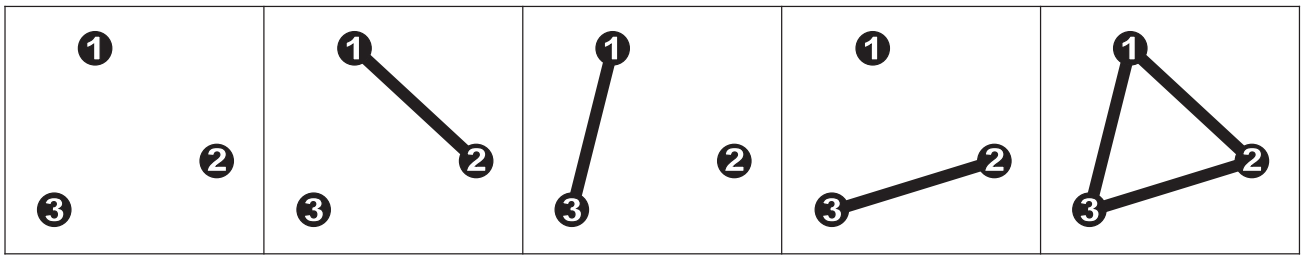
Fellegi and Sunter Method

- Newcombe (1959), Fellegi and Sunter (1969): two databases, all-to-all comparison of records.
- Neyman Pearson hypothesis test with a threshold t.
- If record i and j are above t, then we have a match.
- Otherwise, a non-match.

Fellegi and Sunter Method

- Computationally intractable.
- Transitivity not preserved.
- If 1 matches 2, and 2 matches 3, then 1 does NOT necessarily match 3.

Major limitations and major flaws of this method!



Evaluation Metrics

How do we evaluate performance of a particular record linkage method?

Evaluation Metrics

1. Recall
2. Precision
3. Reduction Ratio
4. Estimated Sample Size
5. Standard Error of Estimated Sample Size
6. Run Time
7. Robustness to Tuning Parameters
8. Do Supervised Methods Overfit the Data

Evaluation Metrics

1. Pairs of data can be linked in both the handmatched training data (which we refer to as "truth") and under the estimated linked data. We refer to this situation as true positives (TP).
2. Pairs of data can be linked under the truth but not linked under the estimate, which are called false negatives (FN).
3. Pairs of data can be not linked under the truth but linked under the estimate, which are called false positives (FP).
4. Pairs of data can be not linked under the truth and also not linked under the estimate, which we refer to as true negatives (TN).

Recall, Precision, F-measure

$$\text{Recall} = \frac{TP}{TP + FN} = 1 - \text{FNR}.$$

Recall, Precision, F-measure

$$\text{Recall} = \frac{TP}{TP + FN} = 1 - \text{FNR}.$$

$$\text{Precision} = \frac{TP}{TP + FP} = 1 - \text{FDR}.$$

$$\text{F-measure} = 2 \times \frac{(precision \times recall)}{(precision + recall)}.$$

Recall, Precision, F-measure

$$\text{Recall} = \frac{TP}{TP + FN} = 1 - \text{FNR}.$$

Recall, Precision, F-measure

$$\text{Recall} = \frac{TP}{TP + FN} = 1 - \text{FNR}.$$

$$\text{Precision} = \frac{TP}{TP + FP} = 1 - \text{FDR}.$$

Recall, Precision, F-measure

$$\text{Recall} = \frac{TP}{TP + FN} = 1 - \text{FNR}.$$

$$\text{Precision} = \frac{TP}{TP + FP} = 1 - \text{FDR}.$$

$$\text{F-measure} = 2 \times \frac{(precision \times recall)}{(precision + recall)}.$$

Recall, Precision, F-measure

$$\text{Recall} = \frac{TP}{TP + FN} = 1 - \text{FNR}.$$

$$\text{Precision} = \frac{TP}{TP + FP} = 1 - \text{FDR}.$$

$$\text{F-measure} = 2 \times \frac{(precision \times recall)}{(precision + recall)}.$$

$$\text{specificity} = \frac{TN}{TN + FP}.$$

$$\text{FPR} = \frac{FP}{TN + FP} = 1 - \text{specificity}.$$

There are issues with using the FPR discussed in Christen (2012) pg. 169 and Steorts (2015) and these should not be utilized for entity resolution problems.

Reduction Ratio

The reduction ratio (RR) measures the relative reduction of the comparison space from the de-duplication or hashing technique.

See Christen (2012), Steorts, Ventura, Sadinle, Fienberg (2014) for a formal definition.

Other Evaluation Metrics

1. Estimated Sample Size
2. Standard Error of Estimated Sample Size
3. Run Time
4. Robustness to Tuning Parameters
5. Do Supervised Methods Overfit the Data

1. The estimated sample size and standard error must be defined for each method, but this is not difficult to do in practice.
2. Any method can be evaluated also for the run time, so one can gauge computationally costs.
3. Robustness of tuning parameters should be explored from a Bayesian and a frequentist perspective.
4. It's also essential to make sure that supervised methods do not overfit the data (see Steorts (2015)).

RLdata500 dataset

Let's look at an example on data that is available from the Record Linkage package in R, where we compare many different methods according to the evaluation metrics that we have laid out.

We will first describe the data set and then compare the following methods in R:

1. blink
2. logistic regression
3. random forests
4. Bayesian adaptive regression trees

RLdata500 dataset

	fname_c1	fname_c2	lname_c1	lname_c2	by	bm	bd
1	CARSTEN	<NA>	MEIER	<NA>	1949	7	22
2	GERD	<NA>	BAUER	<NA>	1968	7	27
3	ROBERT	<NA>	HARTMANN	<NA>	1930	4	30
4	STEFAN	<NA>	WOLFF	<NA>	1957	9	2
5	RALF	<NA>	KRUEGER	<NA>	1966	1	13
6	JUERGEN	<NA>	FRANKE	<NA>	1929	7	4

The `RLdata500` data set consists of 500 records with 10 percent duplication.

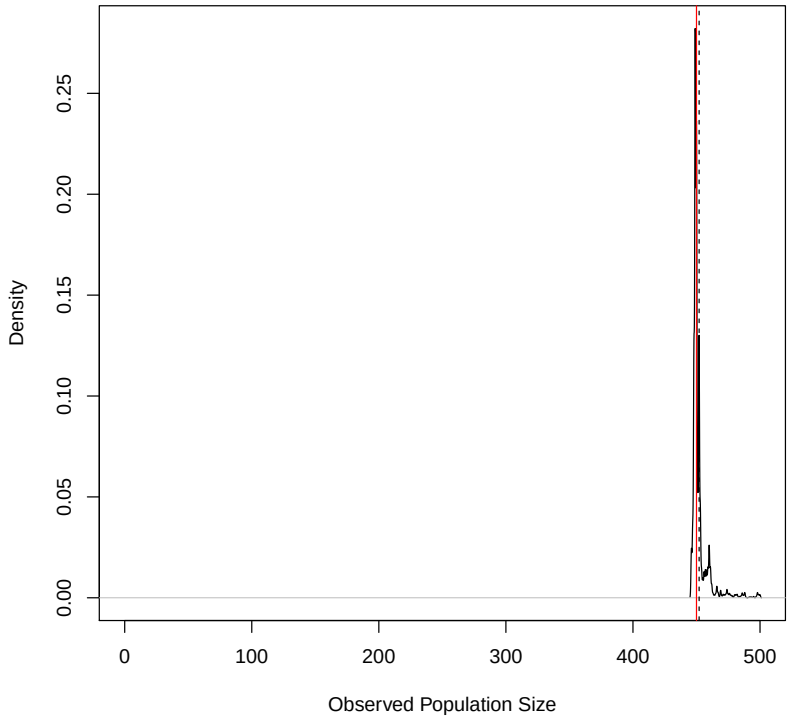


Figure 2: Posterior density for N in simulation study. The FNR and FPR: 0.04 and 0.02.

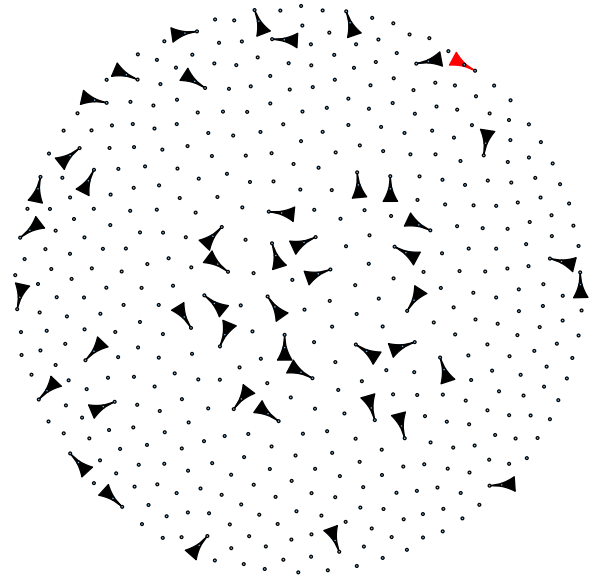


Figure 3: Shared MPMMS graphical representation of simulation study. Only makes one false positive set.

Procedure	FNR	FDR
blink (Steorts (2015))	0.02	0.08
Exact Matching	1	0
Near-Twin	1	0
BART (10% training)	0.10	0.16
BART (20% training)	0.07	0.11
BART (50% training)	0.03	0.04
BART (full data)	0.02	0
Random Forests (10% training)	0.05	0.15
Random Forests (20% training)	0.04	0.09
Random Forests (50% training)	0.02	0.06
Random Forests (full data)	0.04	0.06
Logistic Regression (10% training)	0.09	0.16
Logistic Regression (20% training)	0.06	0.07
Logistic Regression (50% training)	0.02	0.01
Logistic Regression (full data)	0.02	0

Table 1: False negative rate (FNR) and false discovery rate (FDR) for the proposed EB methodology and five other record linkage methods. For the supervised methods, we run 100 iterations of each one and average these such that overfitting is not occurring.

Robustness

How do we make sure a method is robust?

For a semi-supervised method, we want to make sure that it's robust to different choices of the training/test data and any tuning paramater(s).

For probabilistic and Bayesian methods, we want to make sure these methods are robust to choices of hyper-parameters and/or tuning parameters.

Robustness, computational time complexity, and sensitivity analysis can be further explored in Steorts (2015) and Steorts, Hall, Fienberg (2016), Marchant+ (2021).

Thank you! Questions?

Contact: beka@stat.duke.edu
Webpage:resteorts.github.io