(Almost) All of Entity Resolution

Rebecca C. Steorts
joint work with Olivier Binette, PhD Student, Department of Statistical Science Duke University

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).

March 22, 2021
Entity resolution (record linkage or de-duplication) is the process of removing duplicated information from large noisy databases.
Entity resolution (record linkage or de-duplication) is the process of removing duplicated information from large noisy databases.

The purpose of this review is to introduce one to the fundamentals of entity resolution, its applications, and modern developments over the past 61+ years.
What is the purpose of entity resolution?
Record Linkage*

HALBERT L. DUNN, M.D., F.A.P.H.A.

Chief, National Office of Vital Statistics, U. S. Public Health Service,

Each person in the world creates a Book of Life. This Book starts
with birth and ends with death. Its pages are made up of the records of
the principal events in life. Record linkage is the name given to the process
of assembling the pages of this Book into a volume.

The Book has many pages for some

the various important records of a
person’s life.

The two most important pages in the
Book of Life are the first one and the
last one. Consequently, in the process
of record linkage the uniting of the
fact-of-death with the fact-of-birth has
been given a special name, “death
clearance.”
Terminology
The linkage graph

Apple
- Larry Wasserman
- Steve Fienberg
- R.J. Tibshirani
- Ryan Tibshirani

Amazon
- Stephen Fienberg
- Stanley Wasserman
- Ed George
- Robert Tibshirani

Google
- Valerie Ventura
- Steve Feinberg
- Jesse Ventura
- Eddie George
The attribute (full name) of Larry Wasserman
The collection of the record of Larry Wasserman

1014 Murray Hill Avenue
Pittsburgh, PA 15217
412-361-3146
De-duplication, Record linkage, and Entity resolution

- Apple
  - Larry Wasserman
  - Steve Fienberg
  - R.J. Tibshirani
  - Ryan Tibshirani

- Amazon
  - Stephen Fienberg
  - Stanley Wasserman
  - Ed George
  - Robert Tibshirani

- Google
  - Valerie Ventura
  - Steve Feinberg
  - Jesse Ventura
  - Eddie George
Challenges
Challenges of Entity Resolution

- **Costly manual labelling**
  Vast amounts of manually-labelled data are typically required for supervised learning and evaluation.

- **Scalability/computational efficiency**
  Approximations are required to avoid quadratic scaling. Need to ensure impact on accuracy is minimal.

- **Limited treatment of uncertainty**
  Given inherent uncertainties, it’s important to output predictions with confidence regions.

- **Unreliable evaluation**
  Standard evaluation methods return imprecise estimates of performance.
Pipeline Approach
Data Cleaning Pipeline

1. Schema alignment: Map source schemas to a mediated schema.
2. Blocking: Place similar records into partitions.
3. Entity resolution: Identify records that refer to the same entity.
4. Canonicalization: Merge records that refer to the same entity.

Data sources → full_name

given_name + surname

→ full_name

Integrated data
Blocking and Deterministic Record Linkage
Figure: Left: All to all record comparisons. Right: Similar records placed into the same partition.

Blocking places similar records into a partition.
Deterministic record linkage is the most widely used in the literature given that it is

1. scalable
2. rules that can easily be put together
3. and it is easily transferable across disciplines
Deterministic Record Linkage

1. Exact Matching and Off-by-k Matching
2. Scoring Functions (Edit and Jaro-Winkler distances)
3. Putting simple rules together to form complex rules
Case Study from the UNTC


<table>
<thead>
<tr>
<th>Record</th>
<th>Given name</th>
<th>Family name</th>
<th>Year</th>
<th>Month</th>
<th>Day</th>
<th>Municipality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>JOSE</td>
<td>FLORES</td>
<td>1981</td>
<td>1</td>
<td>29</td>
<td>A</td>
</tr>
<tr>
<td>2.</td>
<td>JOSE</td>
<td>FLORES</td>
<td>1981</td>
<td>2</td>
<td>NA</td>
<td>A</td>
</tr>
<tr>
<td>3.</td>
<td>JOSE</td>
<td>FLORES</td>
<td>1981</td>
<td>3</td>
<td>20</td>
<td>A</td>
</tr>
<tr>
<td>4.</td>
<td>JULIAN ANDRES</td>
<td>RAMOS ROJAS</td>
<td>1986</td>
<td>8</td>
<td>5</td>
<td>B</td>
</tr>
<tr>
<td>5.</td>
<td>JILIAM</td>
<td>RMAOS</td>
<td>1986</td>
<td>8</td>
<td>5</td>
<td>B</td>
</tr>
</tbody>
</table>

Table: Duplicated records reproduced from Table 1 of Sadinle (2014). Records 1 – 3 should refer to the same entity. Records 4 – 5 might refer to the same entity. Note that record 5 most likely has OCR errors, where “RMAOS” should be “RAMOS.”
Sadinle (2014) utilized complex rules for a blocking criteria.

Then he applied probabilistic record linkage within each block.

Similar rules are used in other human rights applications by Ball (2006); Sadosky, Shrivastava, Price, Steorts (2015); Chen, Shrivastava, Steorts (2018); and in work established by the Human Rights Data Analysis Group (HRDAG).

While deterministic rules are recommended for intuition or blocking, they are not recommended in place of probabilistic record linkage.

Steorts, Ventura, Sadinle, Fienberg (2014) and Murray (2016) provide reviews on deterministic and probabilistic blocking.
Probabilistic Record Linkage
Record Linkage*

HALBERT L. DUNN, M.D., F.A.P.H.A.

Chief, National Office of Vital Statistics, U. S. Public Health Service,

Halbert L. Dunn (1896-1975):

• “leading figure in establishing a national vital statistics system
  in the United States”
Record linkage is the task of assembling together all important pieces of information which refer to the same individual.
Record Linkage*

HALBERT L. DUNN, M.D., F.A.P.H.A.


EACH person in the world creates a Book of Life. This Book starts with birth and ends with death. Its pages are made up of the records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this Book into a volume.

The Book has many pages for some the various important records of a person's life.

The two most important pages in the Book of Life are the first one and the last one. Consequently, in the process of record linkage the uniting of the fact-of-death with the fact-of-birth has been given a special name, "death clearance."
Automatic Linkage of Vital Records*

Computers can be used to extract "follow-up" statistics of families from files of routine records.

Proposed a probabilistic record linkage method and implemented it on the Datatron 205 computer.
A THEORY FOR RECORD LINKAGE*

Ivan P. Fellegi and Alan B. Sunter
Dominion Bureau of Statistics

A mathematical model is developed to provide a theoretical framework for a computer-oriented solution to the problem of recognizing those records in two files which represent identical persons, objects or events (said to be matched).
Fellegi and Sunter (1969), JASA

The authors formalized Newcombe et al. (1959) in a decision-theoretic framework.

One determines if two records are a match using a likelihood ratio test exceeding a threshold.
There are two methods proposed in this paper.

One method is completely unsupervised, and the other is semi-supervised.

The semi-supervised methods are used in practice for computational reasons.
Both the methods of Newcombe et al. (1959) and Fellegi and Sunter (1969) have led to extensions that are utilized at statistical agencies in order to update our “book keeping” regarding an individual’s book of life.

1. The methods must rely on training data.
2. There is sensitivity to tuning parameters such as the threshold used.
3. Both methods on their own do not scale to large data sets.
Modern Probabilistic Record Linkage
The work of Enamorado et al. (2019) extends Fellegi and Sunter (1969) such that

1. the authors extend Lahiri and Larsen (2005) to incorporate auxiliary information such as population name frequency and migration rates into the merge procedure to conduct post-merge analyses

2. the authors are able to account for uncertainty of the merge process

3. the authors use parallelization and efficient data representations such that they can scale to millions of records

This work has been extended by Enamorado and Steorts (2020) to utilize fastLink as proposal for probabilistic blocking.
Bayesian FS Methods

Sadinle (2014) is a Bayesian Fellegi-Sunter method for de-duplication using a likelihood ratio similar in spirit to Fortini et al. (2001) and Fellegi and Sunter (1969).

The author considers a prior on the matching configuration matrix which imposes *transitive closures* — records are partitioned into groups which are thought to refer to the same entity.

This allows for uncertainty quantification via the posterior distribution.

The author provides corrections and a small labelled set of records for the UNTC data set as well a case study.
Sadinle (2017) extended this work to bipartite record linkage and derived Bayes estimates under a general class of loss functions, providing an alternative to the FS decision rule.

McVeigh et al. (2020) is an extension of Sadinle (2014, 2017) to include probabilistic blocking before the use of Bayesian FS.

The authors greatly improve the speed of the original approach of Sadinle and apply this methodology to a case study on voter registration and historical census records from California.
Semi-supervised and fully supervised methods

Semi-supervised methods use a relatively small amount of manually classified record pairs, known as labeled pairs, to improve upon unsupervised probabilistic record linkage.

Fully supervised methods do not exploit information provided by unlabeled examples; instead they rely on larger amounts of labeled pairs.

1. Training data may come from approximate training sets (using unsupervised ER)
2. Training data may come by manual labelling of data
3. Training data may come from crowdsourcing extensive manual record linkage efforts.

Approximate training sets: Torvik et al. (2015)


Clustering Based Methods
Why graphical Bayesian models?

distributed graphical entity resolution (d-blink)

Joint model for blocking and ER:

1. Scales to large databases using partially collapsed Gibbs sampling
2. Distributed inference/parallel inference is possible
3. Many computational speeds ups are proposed
4. Extensive studies on synthetic and real data sets are given
5. Open source software is provided in both Apache Spark and R.

Open Research Problems and Discussion Questions

1. How should one do fair comparisons and evaluations?
2. How should one create training data sets?
3. How should one find publicly available benchmark data sets and which ones should be used?
4. Should we be aware that entity resolution can be used for massive data collection efforts and evasion of privacy? What can we do as a field regarding this?
5. What privacy guarantees are offered for record linkage?
6. What ethical considerations do I have when working with private data that is sensitive, such as the data that HRDAG provides?
7. What resources are available to me if I want to know more about the field of record linkage?
Questions?
beka@stat.duke.edu
Webpage: resteorts.github.io
Software: https://github.com/orgs/cleanzr/
CRP Behavoir for N increasing and $\alpha$ fixed

As $N$ increases the size of the clusters increases as $O(N) \implies$ Not appropriate for microclustering.
CRP Behaviour for N and $\alpha$ jointly increasing

Can obtain microclustering property by increasing $\alpha$ with $N$ but the resulting model becomes less flexible.

The model collapses onto a one parameter family distribution for the cluster sizes.

This behavior motivates models that naturally incorporate the microclustering property.