# Bayesian Record Linkage

Speaker: Rebecca Steorts

Last update on June 27, 2021

*Review paper* Binette and Steorts (2020)

## 1 Introduction

Usually, in datasets we have some problems of *duplication*: duplicated data has to be removed before inference can be carried out.

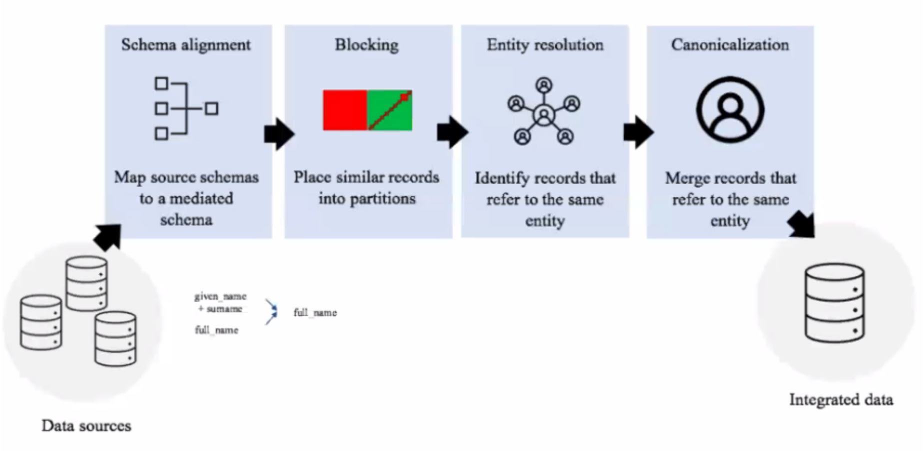
Schema alignment
Blocking
Entity resolution
Canonicalization
Map source schemas to a mediated schema
Place similar records into partitions
Identify records that refer to the same entity
Merge records that refer to the same entity
given_name + surname
full_name
full_name
Data sources
Integrated data

Figure 1: Usual data cleaning pipeline in surveys. Entity resolution is the main focus of this talk.

Entity resolution (also record linkage, data matching, duplicate detection, . . . ) is only one part of the process, which involves merging noisy databases containing duplicate entries and identifying such duplicates Figure 2.
Since entities are real objects (people, businesses, . . . ) we are provided further data in order to automate this process of entity resolution. We do not want to identify the *most representative* node (canonicalization), instead we only want to cluster nodes into the same referenced entities.

### Challenges of entity resolution

a) Lots of manually labeled data are needed for supervised learning.

b) Scalability and computational efficiency, therefore we use approximations to avoid quadratic scaling.

c) It's important to have some credible regions (Bayesian approach) in order to quantify uncertainty.

d) Evaluation methods give noisy estimates of performance, therefore using a Bayesian approach is very important.

Although the first approaches were based on likelihood ratios, modern approaches are based on graphical models (Steorts, Ventura, et al., 2014; Steorts, Hall, et al., 2015). The model proposed in

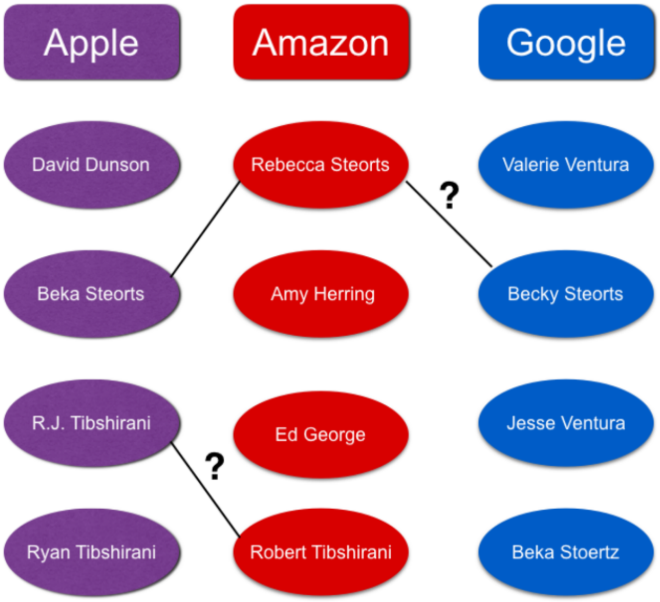

Figure 2: Matching individuals is a problem which involves understanding whether two nodes in a graph are coreferenced or not.

Figure 3 builds off a graphical model in Tancredi and Liseo (2011), and is popular due to its large number of advantages over competing approaches.

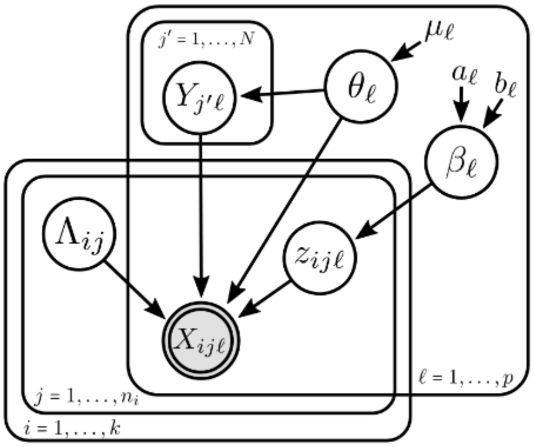

Figure 3: Graphical Entity Resolution model for record linkage.

- Can handle an arbitrary number of databases
- Categorical, textual, missing data, lots of applications (Chen et al., 2017)
- Uncertainty quantification and good properties both theoretical and performance-wise
- Great scalability to high-dimensional datasets.

Moreover, you can mathematically prove that the Bayesian E-R models have tight performance bounds. The model Figure 3 is difficult to scale to millions of data, therefore there is a new proposed scalable joint Bayesian model for blocking and entity resolution (Marchant et al., 2021).

### Microclustering

As the number of data grows we often find that the cluster sizes scales sublinearly, i.e. if $C_n$ is a random partition of the data, the maximum cluster dimension $M_n$ is such that $M_n \sim o_p(n)$ (Zanella et al., 2016; Betancourt et al., 2020). Therefore, standard nonparametric clustering methods such as the Dirichlet Process and Pitman-Yor Process are not appropriate for this type of problem, since they show a power-law behaviour for the resulting random cluster dimensions (Broderick et al., 2011).

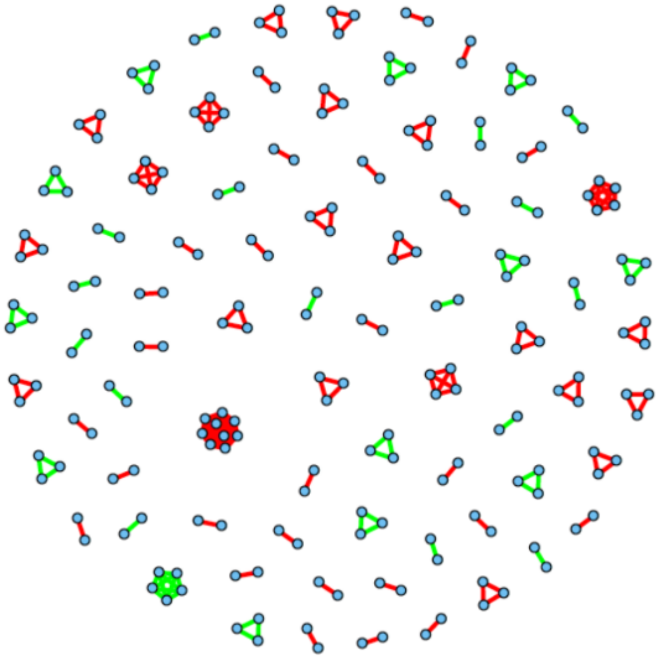

Figure 4: Example of the microclustering behaviour for a longitudinal dataset.

# 2 End-to-end Entity Resolution

We want to join duplicate entities between datasets without having an identifier, instead relying on other information. Usual problems in this literature are

a) Finding models such that their computational cost scales less than $\mathcal{O}(M^2)$.

b) They fit well and have good overall performance.

The literature is usually based on de-duplicating within just one database, and is performed through random forests predictive approaches. Entity resolution is about duplication both within and between datasets.

### Blocking

One often performs blocking, that is, grouping observations in blocks Figure 5, since they usually show some degree of correlation between each other.

Blocking can be performed via

- Deterministic blocking method, e.g. block by sex/age/. . . , which usually end up as not being very robust.
- Probabilistic blocking methods, such as locally-sensitive hashing which are based on locality measures (Jaccard, cosine metric, . . . ).

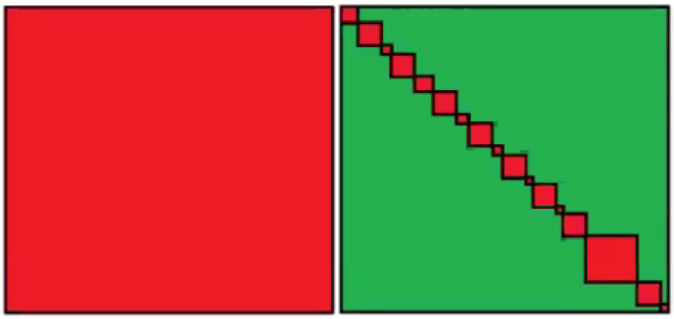

Figure 5: Result of blocking observations in a dataset.

The main contribution of (Marchant et al., 2021) is developing a graphical model in order to jointly perform blocking and entity resolution, so that errors during the blocking procedure can be propagated forward in the pipeline.

## 2.1 Main contribution

The graphical Bayesian ER introduced in Marchant et al. (2021) is able to do so, by introducing conditional independence between latent entities, which translates into a blocking mechanism for records given the value of the latent variable. Then, similar entities are grouped via kdtrees, onto which a partially-collapsed Gibbs sampling algorithm is applied in the context of a distributed computing environment.

Some computational tricks are introduced:

a) Sub-quadratic algorithm based on indexing

b) Truncation of the attribute similarities.

c) Perturbation sampling algorithm which relies on the Vose-Alias method to efficiently sample from a discrete probability distribution (Walker, 1974; Vose, 1991).

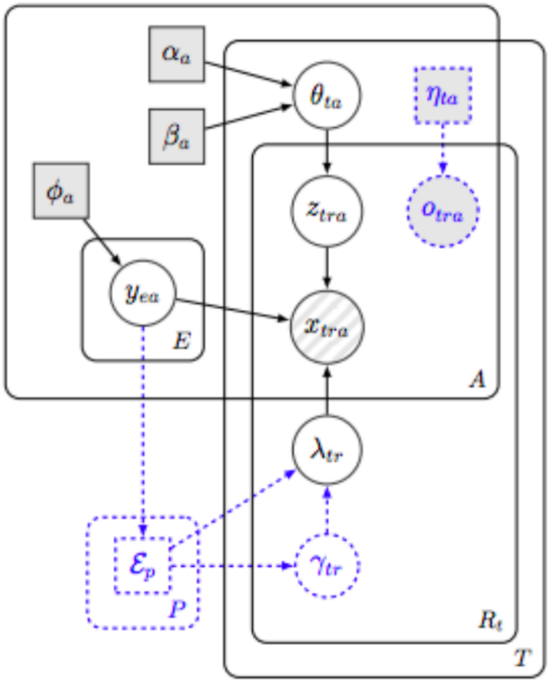

Figure 6: Graphical representation of the dblink model (Marchant et al., 2021). Plates represent entities $E$, attributes $A$, records $R_t$, and tables $T$.

The distributed computing environment lets it compute the summary statistics conditional on the parameter updates. Two computing bottlenecks can be found in

› The *linkage structure update*: it can be resolved by maintaining and index of entity attributes → entities and entities → linked records, in order to prune candidate links for a record based on a similarity score thresholding.

› The *entity attribute update*: we can express it as a two-component perturbation mixed model.

Using a fully parallel MCMC sampling method such as Wasserstein Subset Posterior Barycenters (Srivastava et al., 2015) is not satisfactory, since it does not fully characterizes the joint model.

# References

Betancourt, B. et al. (2020). "Random Partition Models for Microclustering Tasks". In: *Journal of the American Statistical Association* 0.0, pp. 1–13.

Binette, O. and Steorts, R. C. (2020). *(Almost) All of Entity Resolution*. arXiv: 2008.04443 [cs, stat].

Broderick, T. et al. (2011). *Beta Processes, Stick-Breaking, and Power Laws*. arXiv: 1106.0539 [stat].

Chen, B. et al. (2017). *Unique Entity Estimation with Application to the Syrian Conflict*. arXiv: 1710.02690 [cs, stat].

Marchant, N. G. et al. (2021). "D-blink: Distributed End-to-End Bayesian Entity Resolution". In: *Journal of Computational and Graphical Statistics* 30.2, pp. 406–421.

Srivastava, S. et al. (2015). "WASP: Scalable Bayes via barycenters of subset posteriors". In: *Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. PMLR, pp. 912–920.

Steorts, R. C., Hall, R., et al. (2015). *A Bayesian Approach to Graphical Record Linkage and De-Duplication*. arXiv: 1312.4645 [stat].

Steorts, R. C., Ventura, S. L., et al. (2014). *A Comparison of Blocking Methods for Record Linkage*. arXiv: 1407.3191 [cs, stat].

Tancredi, A. and Liseo, B. (2011). "A hierarchical Bayesian approach to record linkage and population size problems". In: *The Annals of Applied Statistics* 5 (2B), pp. 1553–1585.

Vose, M. (1991). "A linear algorithm for generating random numbers with a given distribution". In: *IEEE Transactions on Software Engineering* 17.9, pp. 972–975.

Walker, A. J. (1974). "New fast method for generating discrete random numbers with arbitrary frequency distributions". In: *Electronics Letters* 10.8, pp. 127–128.

Zanella, G. et al. (2016). *Flexible Models for Microclustering with Application to Entity Resolution*. arXiv: 1610.09780 [math, stat].