End-to-End Bayesian Entity Resolution

Rebecca C. Steorts

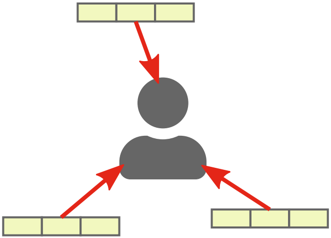
Department of Statistical Science, affiliated faculty in
Computer Science, Biostatistics and Bioinformatics, the
information initiative at Duke (iiD) and
the Social Science Research Institute (SSRI)
Duke University and U.S. Census Bureau

joint work with Neil Marchant, Ben Rubinstein (Melbourne),
Andee Kaplan (CSU), and Daniel Elazar (ABS)

June 29, 2020

Entity resolution

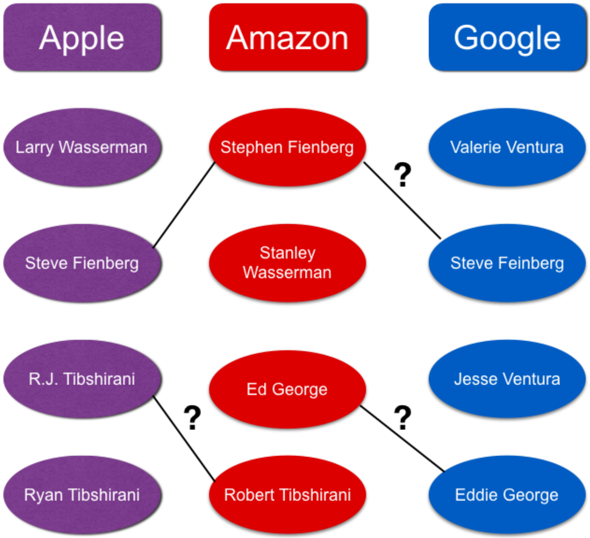
Identifying records across and/or within data sources that refer to the same entities



Also known as:

- record linkage
- data matching
- de-duplication
- data integration

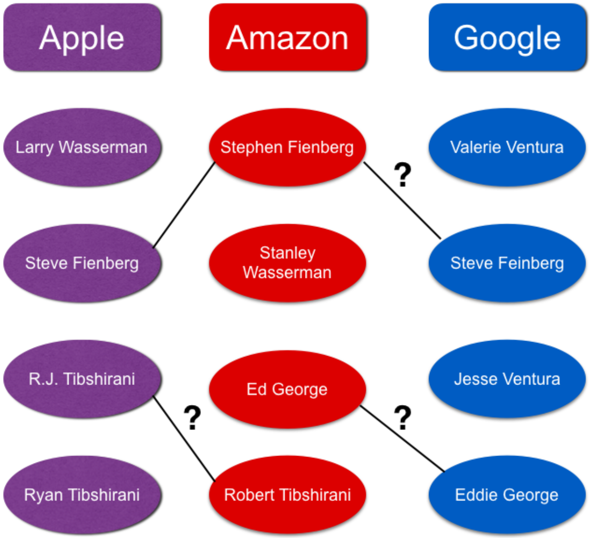
The entity resolution graph



The node of Larry Wasserman

The node of Larry Wasserman

The entity resolution graph



Steve Feinberg

240 Collins Dr
Pittsburgh PA 15235

50-54

412-793-3313

Stephen Fienberg

537 N Neville St Apt 5d
Pittsburgh PA 15213

65+

412-683-5599

Steve Feinberg

240 Collins Dr
Pittsburgh PA 15235

50-54

412-793-3313

Stephen Fienberg

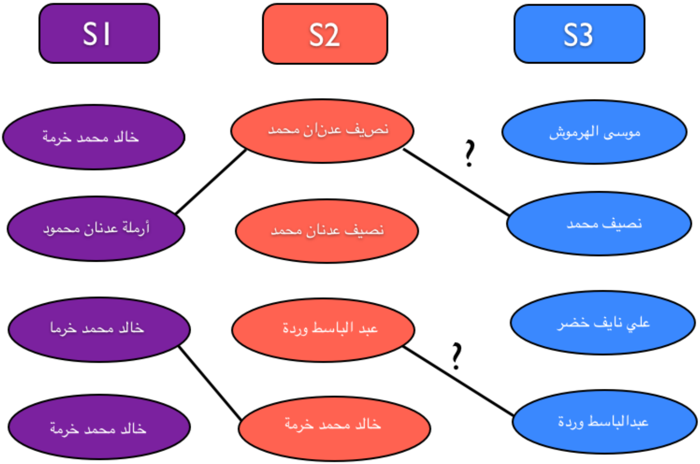
537 N Neville St Apt 5d
Pittsburgh PA 15213

65+

412-683-5599

These are clearly not the *same* Steve Fienberg!

Syrian Civil War



Entity Resolution

Why is entity resolution difficult?

Goals of Entity Resolution

Suppose that we have a total of N records in k databases.

1. We seek models that are much less than $O(N^k)$.
2. We seek models that are reliable, accurate, fit the data well, and account for the uncertainty of the model.
3. We seek models and algorithms to handle unbalanced data (containing duplications).

Existing ER methods

1. deterministic linking
2. probabilistic linking (Fellegi Sunter, random forests, deep learning)
3. Bayesian Fellegi Sunter

Existing ER methods

1. deterministic linking
2. probabilistic linking (Fellegi Sunter, random forests, deep learning)
3. Bayesian Fellegi Sunter

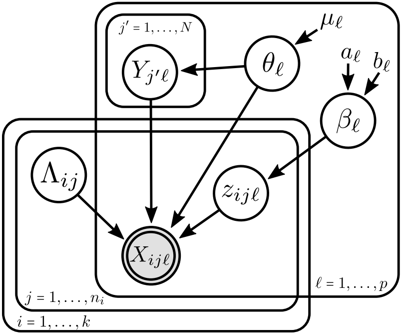
Drawbacks:

- subjectivity in setting the decision threshold
- lack of uncertainty quantification
- require training data

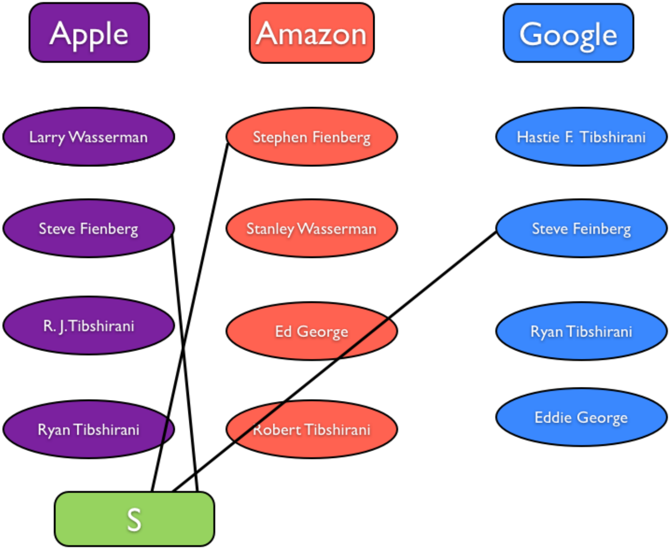
[Fellegi and Sunter (1969), Ventura et al. (2014), Christen (2012), Dong and Shrivastava (2015), Belin and Rubin (1995), Gutman et al. (2013), McVeigh et al. (2020), Sadinle (2014+)].

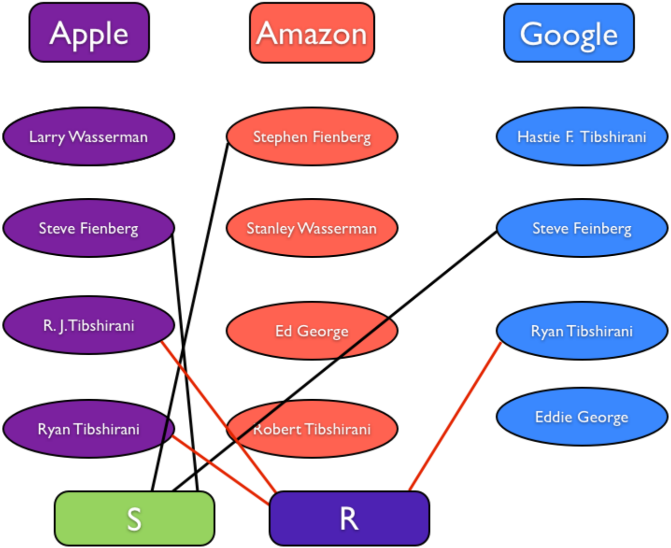
Graphical Bayesian ER

Builds off Copas and Hilton (2011), Tancredi and Liseo (2011).



[**RCS**, Hall, Fienberg (2014, 2016); **RCS** (2015), Zanella, et al. (2016), **RCS** et al. (2017), (2018), Tancredi et al. (2019), Betancourt et al. (2020)].


Apple
Amazon
Google
Larry Wasserman
Stephen Fienberg
Hastie F. Tibshirani
Steve Fienberg
Stanley Wasserman
Steve Feinberg
R. J.Tibshirani
Ed George
Ryan Tibshirani
Ryan Tibshirani
Robert Tibshirani
Eddie George
S


Apple
Amazon
Google
Larry Wasserman
Stephen Fienberg
Hastie F. Tibshirani
Steve Fienberg
Stanley Wasserman
Steve Feinberg
R. J. Tibshirani
Ed George
Ryan Tibshirani
Ryan Tibshirani
Robert Tibshirani
Eddie George
S
R

Our Goal

Scaling Bayesian ER methods to millions of records without sacrificing accuracy and crucially giving uncertainty of the ER task

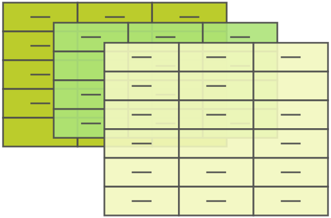
Our Solution

We propose a scalable joint (Bayesian) model for blocking and performing entity resolution, where the error from this joint task is exactly measured.

Problem setup

Key assumptions:

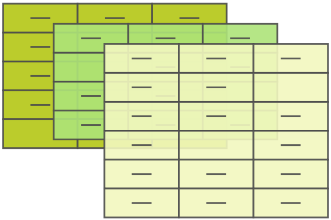
- multiple tables/sources
- duplicates within and across tables
- attributes are aligned
- attributes are discrete
- some missing values
- no ground truth (unsupervised)



Problem setup

Key assumptions:

- multiple tables/sources
- duplicates within and across tables
- attributes are aligned
- attributes are discrete
- some missing values
- no ground truth (unsupervised)

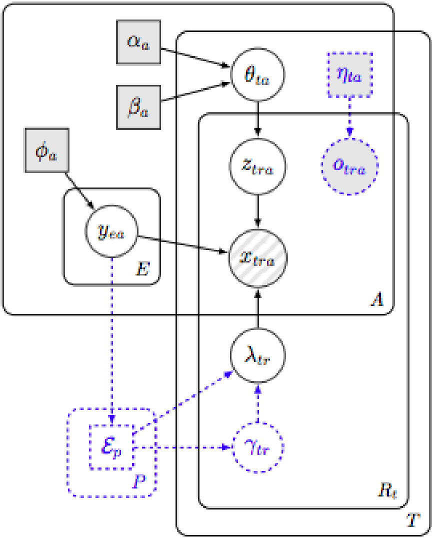


Output: approximate posterior distribution over the linkage structure

1. We propose a joint Bayesian model for blocking (latent entities) and entity resolution.
2. We propose blocks (auxiliary partitions) that induce conditional independencies between the latent entities. This enables distributed inference at the partition-level.
3. The blocking function (responsible for partitioning the entities) groups similar entities together while achieving well-balanced partitions.
4. Application of partially-collapsed Gibbs sampling in the context of distributed computing.
5. Improving computational efficiency:
   a) Sub-quadratic algorithm for updating links based on indexing.
   b) Truncation of the attribute similarities.
   c) Perturbation sampling algorithm for updating the entity attributes, which relies on the Vose-Alias method.

Marchant, **RCS**, Kaplan, Rubinstein, and Elazar (2020).

dblink



α_a
β_a
ϕ_a
θ_{ta}
η_{ta}
z_{tra}
o_{tra}
y_{ea}
x_{tra}
E
A
λ_{tr}
$\mathcal{E}_p$
γ_{tr}
P
R_t
T

Distributed Markov chain Monte Carlo

Since the posterior for the linkage structure $p(\Lambda|X)$ is not tractable, we resort to approximate inference.

Distributed Markov chain Monte Carlo

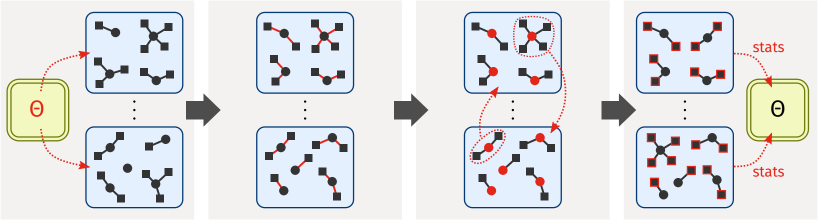
Since the posterior for the linkage structure $p(\Lambda|X)$ is not tractable, we resort to approximate inference.

We propose an MCMC algorithm based on the partially-collapsed Gibbs framework (van Dyk and Park, 2008):

- regular Gibbs updates for the distortion probabilities θ_{ta}, distortion indicators z_{tra} and links λ_{tr}
- "marginalization" and "trimming" are applied to jointly update the entity attributes y_{ea} and the partition assignments for the linked records
- order of the updates is important (to preserve the stationary distribution)

Distributed Markov chain Monte Carlo

Step 1	Step 2	Step 3	Step 4



Update Θ on the master and broadcast to the workers.	Update Λ on the workers. Records may only link to entities within their assigned partitions.	Update Y and Γ on the workers. Move the entities and records to their newly-assigned partitions.	Update Z, then calculate summary stats on the workers. Broadcast to the master.

Tricks for speeding up inference

Two main bottlenecks:

1. linkage structure update $\mathcal{O}(\#\text{ records} \times \#\text{ entities})$
2. entity attribute update $\mathcal{O}(\#\text{ entities} \times \text{domain size})$

Tricks for speeding up inference

Two main bottlenecks:

1. linkage structure update $\mathcal{O}(\#\text{ records} \times \#\text{ entities})$
2. entity attribute update $\mathcal{O}(\#\text{ entities} \times \text{domain size})$

Solutions:

1. Indexing: Maintain indices from "entity attributes $\rightarrow$ entities" and "entities $\rightarrow$ linked records." This allows us to prune candidate links for a record
2. Thresholding similarity scores
3. Express the distribution for the entity attribute update as a two-component perturbation mixture model

Experiments

- `ABSEmployee`. A synthetic data set used internally for linkage experiments by the ABS.
- `NCVR`. Two snapshots from the North Carolina Voter Registration database taken two months apart.
- `NLTCS`. A subset of the National Long-Term Care Survey comprising the 1982, 1989 and 1994 waves.
- `SHIW0810`. A subset from the Bank of Italy's Survey on Household Income and Wealth comprising the 2008 and 2010 waves.
- `RLdata10000`. A synthetic data set provided with the `RecordLinkage` R package.

Experiments

- Implemented `d-blink` and baselines in Apache Spark
- Ran experiments on a local server and Amazon EMR
- (Mostly) used a sample size of 10^3 after burnin (of 10^3 iterations) and thinning (keeping every 10th iteration)
- 3 real and 2 synthetic data sets

Experiments

- Implemented `d-blink` and baselines in Apache Spark
- Ran experiments on a local server and Amazon EMR
- (Mostly) used a sample size of 10^3 after burnin (of 10^3 iterations) and thinning (keeping every 10th iteration)
- 3 real and 2 synthetic data sets

Data set	# records	# tables	# entities	# attributes	
				categorical	string
⋆ `ABSEmployee`	600,000	3	400,000	4	0
`NCVR`	448,134	2	296,433	3	3
`NLTCS`	57,077	3	34,945	6	0
`SHIW0810`	39,743	2	28,584	8	0
⋆ `RLdata10000`	10,000	1	9,000	2	3

Table: Assessment of the pairwise linkage performance for dblink and FS method as our baseline. We note that FS is supervised and does not propagate the entity resolution error exactly compared to dblink.[1]

Data set	Method	Pairwise measure		
		Precision	Recall	F1-score
ABSEmployee	dblink	**0.9943**	**0.8867**	**0.9374**
	Fellegi-Sunter (100)	0.9964	0.9510	0.9736
	Fellegi-Sunter (10)	0.4321	0.6034	0.9736
NCVR	dblink	**0.9179**	**0.9654**	**0.9411**
	Fellegi-Sunter (100)	0.8989	0.9974	0.9456
	Fellegi-Sunter (10)	0.8989	0.9974	0.9456
NLTCS	dblink	**0.8363**	**0.9102**	**0.8717**
	Fellegi-Sunter (100)	0.7969	0.9959	0.8853
	Fellegi-Sunter (10)	0.1902	0.9999	0.3196

[1] Comparisons to other semi-supervised methods are the same.

Posterior Bias Plot

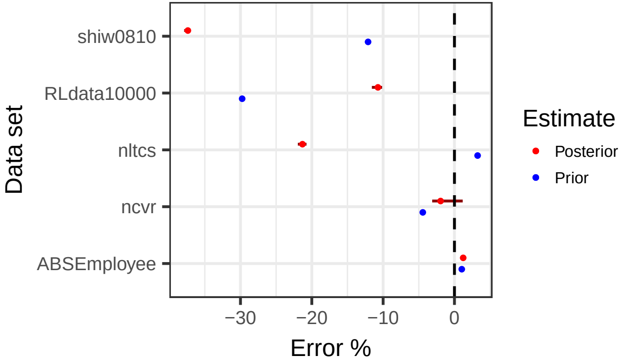
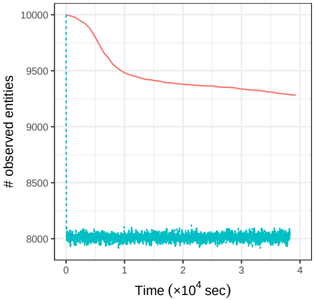
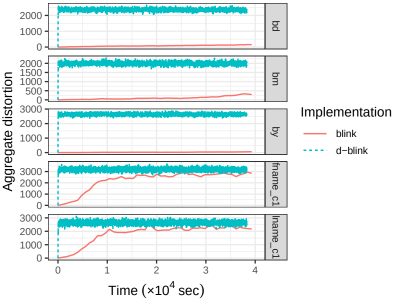


Figure: Error in the posterior and prior estimates for the number of observed entities for d-blink. The results show that the posterior estimate is very sharp and typically underestimates the true number, which is consistent with **RCS**, Hall, Fienberg (2016).

Convergence of d-blink versus blink

We examined the rate of convergence of d-blink versus blink on RLdata10000 without partitioning.





d-blink converges rapidly, however blink fails to reach the equilibrium distribution within 11 hours.

Ongoing work

1. Developing a general set of Bayesian ER models that are non-parametric and allow for more automated tuning of any parameters.
2. Allowing this model to be flexible to names that are not English (Hispanic, Arabic, etc).
3. Developing a parallelized algorithm for this model.
4. Integration of this into d-blink, which is quite extensive.
5. Pushing the limits of scalability.
6. Models for structured and unstructured databases in an unsupervised manner.

Questions?

Contact: beka@stat.duke.edu

This work has been supported by NSF CAREER and the Sloan Foundation, and the ideas of this paper are of the authors and not of the granting organization.

https://github.com/resteorts/record-linkage-tutorial
https://arxiv.org/abs/1909.06039
https://github.com/cleanzr/dblink