d-blink: Distributed 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

July 9, 2019

Motivation

1. Enumerating a census.
2. Enumerating those that have died in a conflict (such as Syria).
3. Predicting those in poverty in small regions from survey data.
4. Predicting results of elections from voter registration data.

Each task may contain duplicated information, which is problematic for the underlying task at hand.

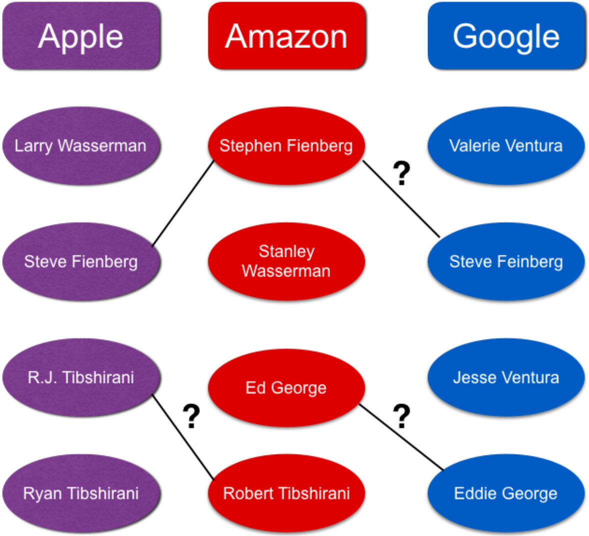
Scalable Entity Resolution

How does one scale Bayesian entity resolution methods in real time?

Entity Resolution

Entity resolution (record linkage or de-duplication) is the process of merging together noisy databases to remove duplicate entities.

The entity resolution graph

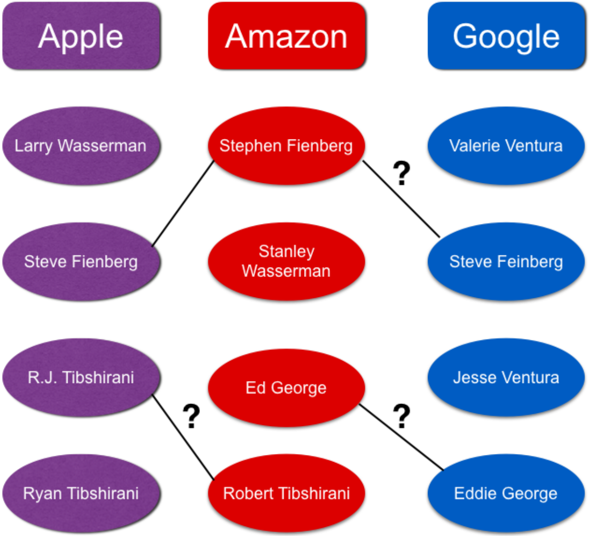


The node of Larry Wasserman

The node of Larry Wasserman

1014 Murray Hill Avenue
Pittsburgh, PA 15217
412-361-3146

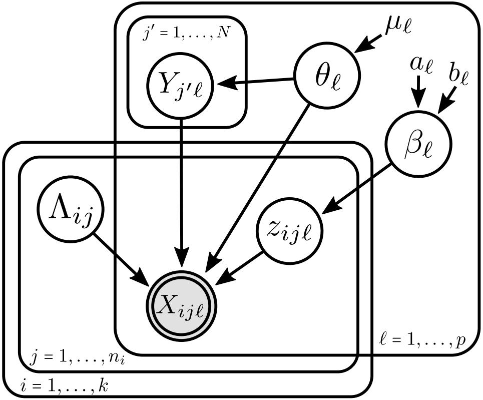
The entity resolution graph



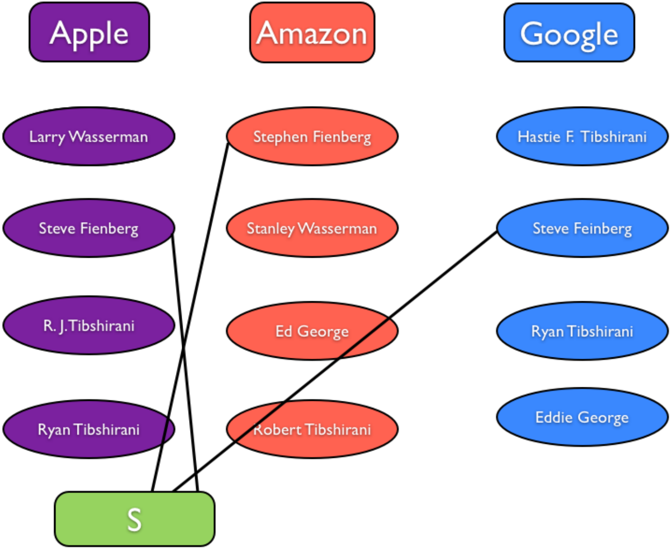
Scalable Bayesian entity resolution

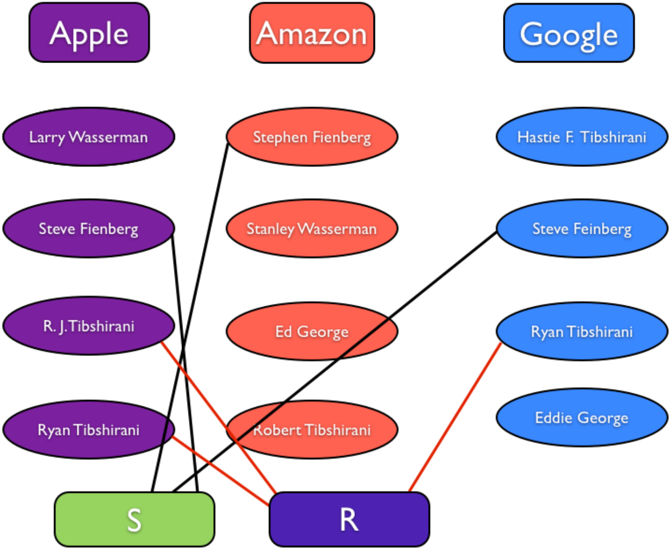
How does one scale Bayesian entity resolution in real time?

Graphical entity resolution



[Copas and Hilton (1990), Tancredi and Liseo (2011), **RCS**, Hall, Fienberg (2014, 2016); Sadinle (2014, 2016); **RCS** (2015)].


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

Why this approach?

1. It models the entities as latent variables, which ensures transitivity and provides a merged representation of linked records.
2. It supports multiple tables (databases).
3. The distortion process can be informed by a string similarity/distance function.
4. Tight performance bounds for entity resolution can be obtained.

[**RCS** (2015); **RCS**, Barnes, and Neiswanger (2017)]
`blink` software can be found on `CRAN` and `github`

Our contribution

We develop the first approach to scaling Bayesian models for entity resolution by integrating the blocking stage directly into the model and by developing sampling amenable to distributed computing.

Our key contributions include:

1. Incorporating auxiliary partitions in the model that induce conditional independencies between the entities. This enables distributed inference at the partition-level, while crucially preserving the marginal posterior of the original model.
2. A partition function (responsible for partitioning the entities), which groups similar entities together while achieving well-balanced partitions.
3. Application of partially-collapsed Gibbs sampling in the context of distributed computing.
4. Incorporating missing values into the modeling framework.
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, Rubenstein, and Elazar (2019), Under Review.

Notation

$t \in \{1 \ldots T\}$	index over tables
$r \in \{1 \ldots R_t\}$	index over records in table t
$e \in \{1 \ldots E\}$	index over entities
$p \in \{1 \ldots P\}$	index over partitions
$a \in \{1 \ldots A\}$	index over attributes
$v \in \{1 \ldots \lvert\mathcal{V}_a\rvert\}$	index over domain of attribute a
$R = \sum_t R_t$	total number of records across all tables
x_{tra}	value of attribute a for record r in table t
z_{tra}	distortion indicator for x_{tra}
o_{tra}	observed/missing indicator for x_{tra}
y_{ea}	value of attribute a for entity e
γ_{tr}	partition assignment for record r in table t
λ_{tr}	entity assignment for record r in table t
θ_{ta}	prob. that attribute a in table t is distorted
α_a, β_a	distortion hyperparameters for attribute a
η_{ta}	prob. that attribute a in table t is observed
$\mathcal{V}_a$	domain of attribute a (indexed by v)
$\phi_a(\cdot)$	distribution over domain of attribute a
$\mathtt{sim}_a(\cdot, \cdot)$	similarity measure for attribute a
$\mathcal{E}_e$	set of records assigned to entity e
$\mathcal{P}_p$	set of entities assigned to partition p
$\mathtt{PartFn}(\cdot)$	partition assignment function

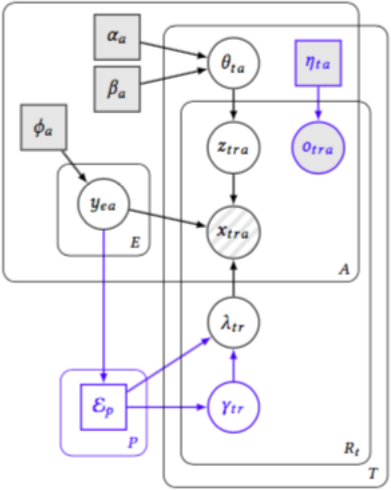


Figure: Circular nodes denote random variables, square nodes denote deterministic variables, (un)shaded nodes denote (un)observed variables, and arrows denote conditional dependence.

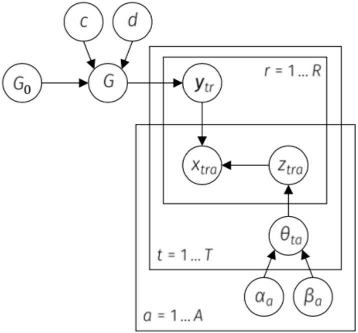
Remainder of the talk

1. Present a simplified version of the model and tricks for performing inference.
2. Why? To provide more intuition to users and practitioners.
3. Present experiments/demos using the generalized model.

A demo of the software will be presented at the end of the talk in Apache Spark.

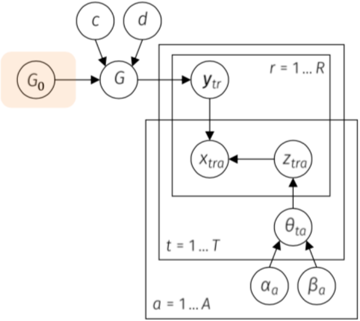
Overview of d-blink

1. Entity attributes
   - Use the empirical distribution function
   - Assume attributes are independent
2. Links from entities to attributes
3. Record attributes
   - Hit and miss distortion prior
   - When distorted draw from attribute domain based on similarity to non-distorted value



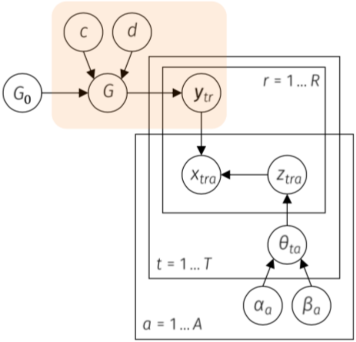
Overview of d-blink

1. Entity attributes
   - Use the empirical distribution function
   - Assume attributes are independent
2. Links from entities to attributes
3. Record attributes
   - Hit and miss distortion prior
   - When distorted draw from attribute domain based on similarity to non-distorted value



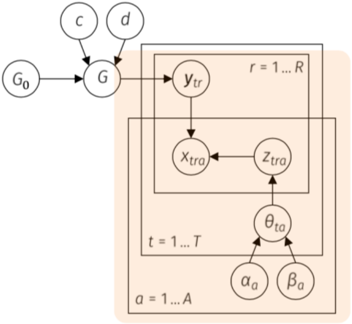
Overview of d-blink

1. Entity attributes
   - Use the empirical distribution function
   - Assume attributes are independent
2. Links from entities to attributes
3. Record attributes
   - Hit and miss distortion prior
   - When distorted draw from attribute domain based on similarity to non-distorted value



Overview of d-blink

1. Entity attributes
   - Use the empirical distribution function
   - Assume attributes are independent
2. Links from entities to attributes
3. Record attributes
   - Hit and miss distortion prior
   - When distorted draw from attribute domain based on similarity to non-distorted value

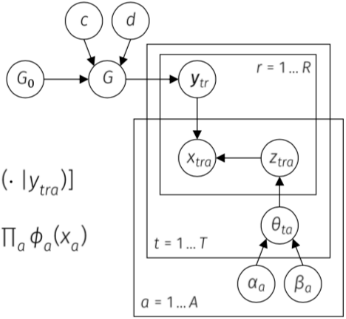


d-blink model

$$
\begin{aligned}
G &\sim \mathrm{PYP}[G_0; c, d] \\
\boldsymbol{y}_{tr} \mid G &\sim G \\
\theta_{ta} &\sim \mathrm{Beta}[\alpha_a, \beta_a] \\
z_{tra} \mid \theta_{ta} &\sim \mathrm{Bernoulli}(\theta_{ta}) \\
x_{tra} \mid z_{tra}, y_{tra} &\sim (1 - z_{tra})\delta_{x_{tra}} + z_{tra}\,\mathrm{Discrete}[\psi(\cdot \mid y_{tra})]
\end{aligned}
$$

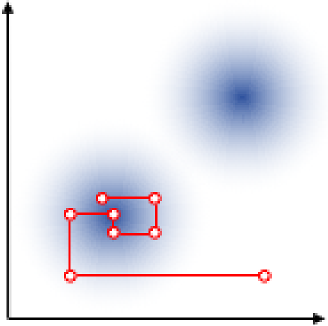
where the density for G_0 is given by $p_0\ \mathbf{x}\ = \prod_a \phi_a(x_a)$

and $\psi\ x|y\ \propto \phi_a(x)e^{\mathrm{sim}_a(y,x)}$



Gibbs sampler

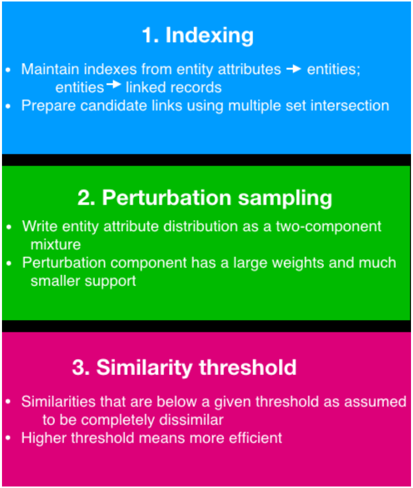
1. Reduce the problem to a sequence of low-dimensional simulations:
   - Conditional distributions must be known and easy to sample from
2. Details for this model
   - Partially-collapse the distortion indicators to improve mixing
   - Need to introduce auxiliary variables to update the hyper-priors



Gibbs sampling tricks

Naive approach scales pooly:

1. Linkage structure update: O(# entities ×# records)
2. Entity attribute update: O(# entities × domain size)



1. Indexing

- Maintain indexes from entity attributes → entities; entities → linked records
- Prepare candidate links using multiple set intersection

2. Perturbation sampling

- Write entity attribute distribution as a two-component mixture
- Perturbation component has a large weights and much smaller support

3. Similarity threshold

- Similarities that are below a given threshold as assumed to be completely dissimilar
- Higher threshold means more efficient

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.

Results

The proposed method is applied to three real applications, with comparisons to standards in the literature.

Table: Summary of the data sets used in the experiments. Data sets marked with a '⋆' are synthetic.

Data set	# records (M)	# files (K)	# entities	# attributes (L)	
				cat.	str.
⋆ ABSEmployee	594,619	3	372,053	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.

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
SHIW0810	dblink	**0.2529**	**0.5378**	**0.3440**
	Fellegi-Sunter (100)	0.1263	0.8480	0.2198
	Fellegi-Sunter (10)	0.0947	0.9244	0.1719
RLdata10000	dblink	**0.6310**	**0.9970**	**0.7729**
	Fellegi-Sunter (100)	0.9153	0.9940	0.9530
	Fellegi-Sunter (10)	0.1706	1.0000	0.2915

Error in number of observed entities

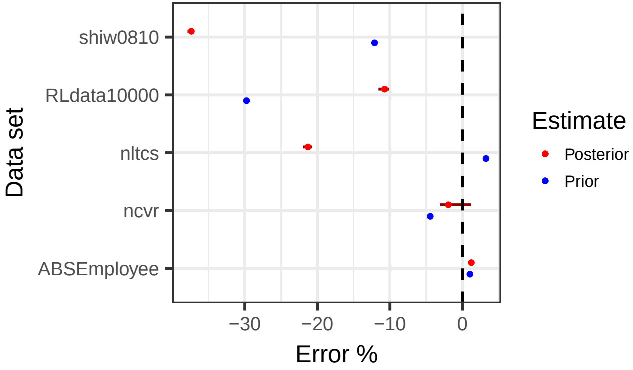


Figure: Error in the posterior and prior estimates for the number of observed entities for dblink. The results show that the posterior estimate is very sharp and typically underestimates the true number.

Super-linear speed-up

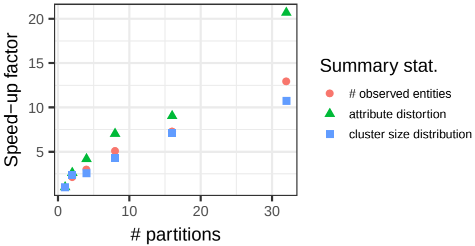


Figure: Efficiency of d-blink as a function of partition size P and summary statistic of interest. The speed-up measures the ESS rate relative to the ESS rate for P = 1 (no partitioning) for the NLTCS data set.

Demo of dblink

1. I will now give a quick demo of the dblink in Apache Spark on the `RLdata500` data set.
2. dblink once available will be public at `https://github.com/ngmarchant/dblink`.
3. The guide for running the package can be found at `https://github.com/ngmarchant/dblink/blob/master/docs/guide.md`

Takeaways

1. Our methods are very scalable and the first of our knowledge for Bayesian entity resolution.
2. The Bayesian approach allows for a full posterior and exact error propagation into downstream tasks (like regression).
3. Our methods perform as well if not better than state-of-the-art methods in practice.
4. With more computing resources, we can scale to larger data sets, which we're currently investigating.

Thank you!

Questions?
beka@stat.duke.edu
resteorts.github.io

Thank you to the National Science Foundation for NSF CAREER Microclustering and NSF Big Data Privacy. The views in this talk are of the authors alone and not of the funding organization.