End to End Entity Resolution

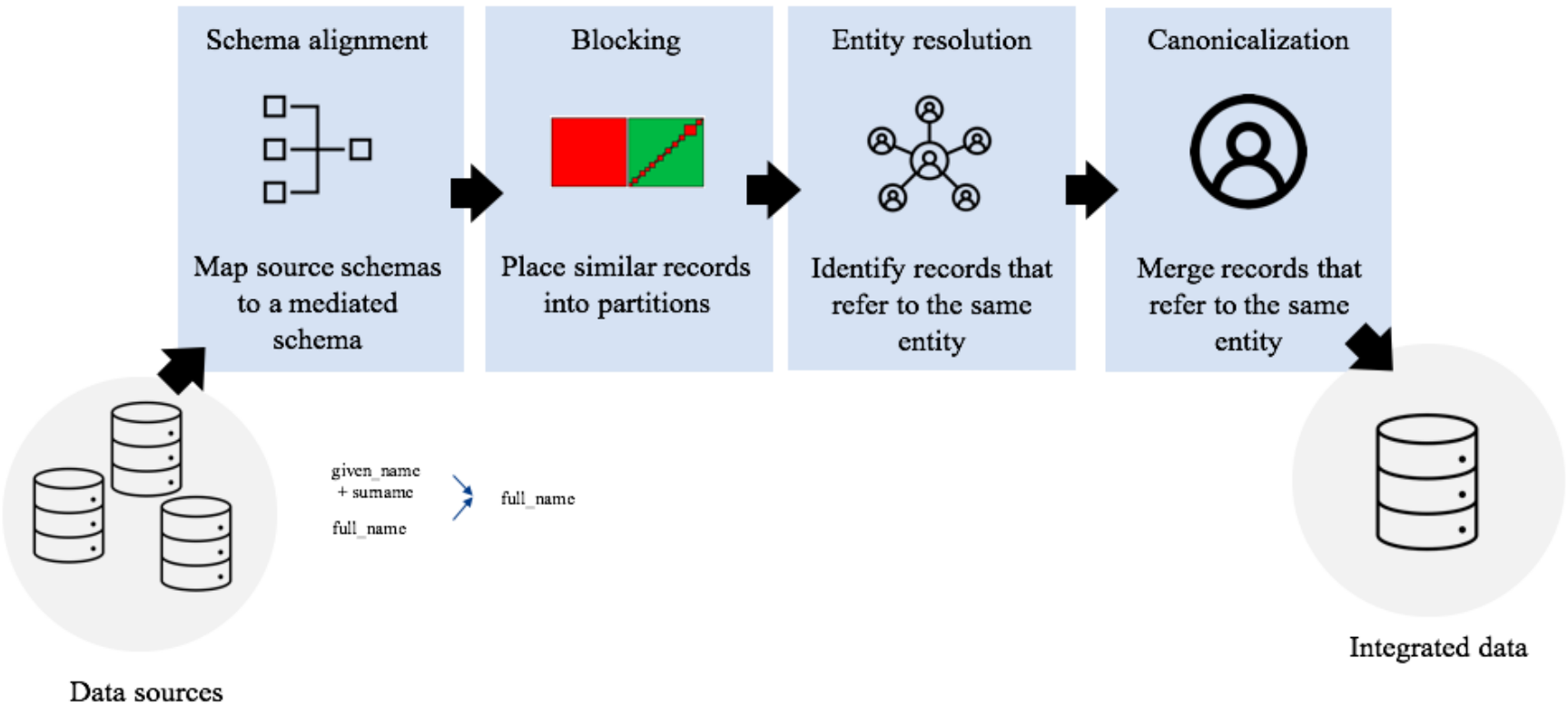
Rebecca C. Steorts

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

Data Cleaning Pipeline



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:

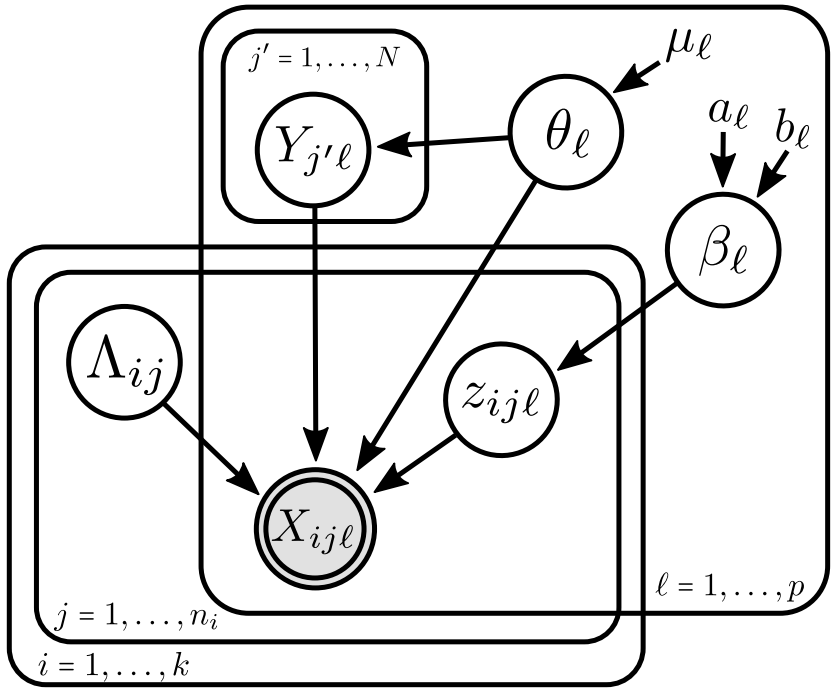
- subjectivity in setting the decision threshold
- lack of uncertainty quantification
- require training data

[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), Sadinle (2017), Sadinle (2018)].

Graphical Entity Resolution

Graphical Bayesian ER

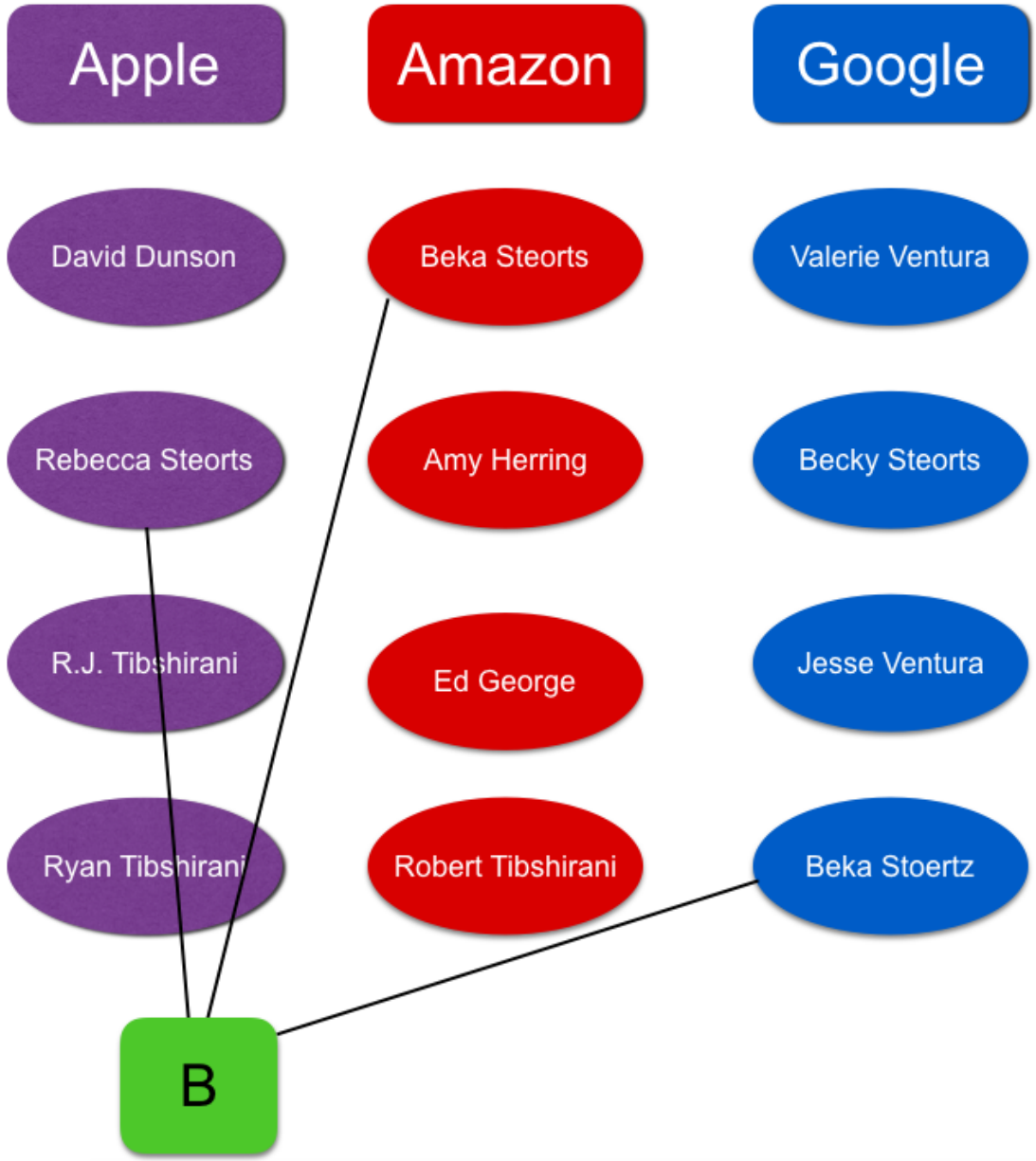
Builds off Copas and Hilton (2001), Tancredi and Liseo (2011).

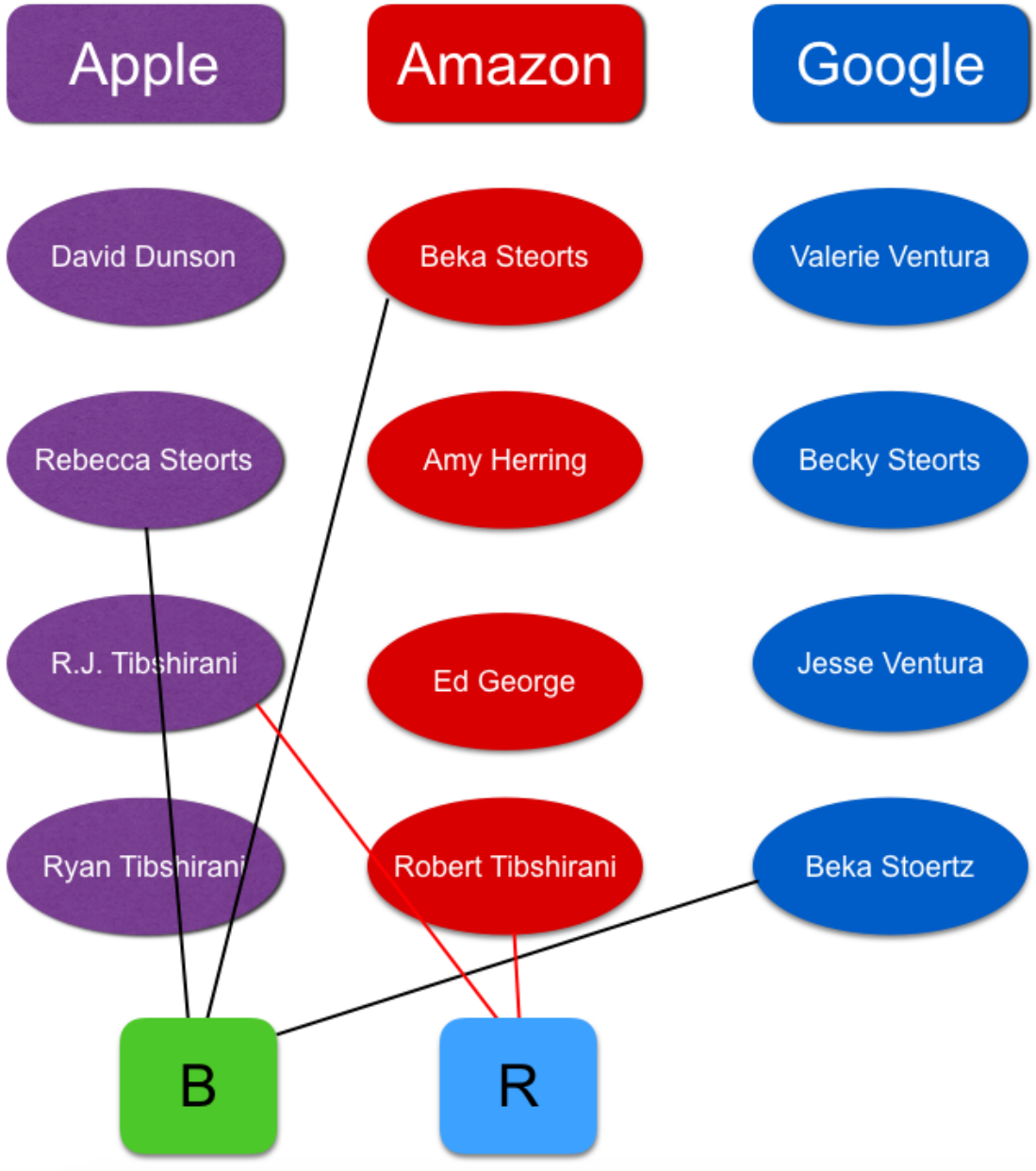


[Steorts+ 2014 AISTATS, Steorts+ 2016 JASA, Steorts 2015 BA]

Why Graphical Bayesian ER

1. Handles any number of databases simultaneously
2. Handles both categorical and textual data
3. Handles missing data
4. Uncertainty quantification is natural
5. Transitive closures are nearly free
6. Has sound theoretical properties
7. Can scale to databases that contain millions of records
8. Generalizes to a wide variety of applications
9. Has equivalent or better performance than alternatives
10. All software is open source and freely available to non-profits


Apple
Amazon
Google
David Dunson
Beka Steorts
Valerie Ventura
Rebecca Steorts
Amy Herring
Becky Steorts
R.J. Tibshirani
Ed George
Jesse Ventura
Ryan Tibshirani
Robert Tibshirani
Beka Stoertz
B


Apple
Amazon
Google
David Dunson
Beka Steorts
Valerie Ventura
Rebecca Steorts
Amy Herring
Becky Steorts
R.J. Tibshirani
Ed George
Jesse Ventura
Ryan Tibshirani
Robert Tibshirani
Beka Stoertz
B
R

Our Goal

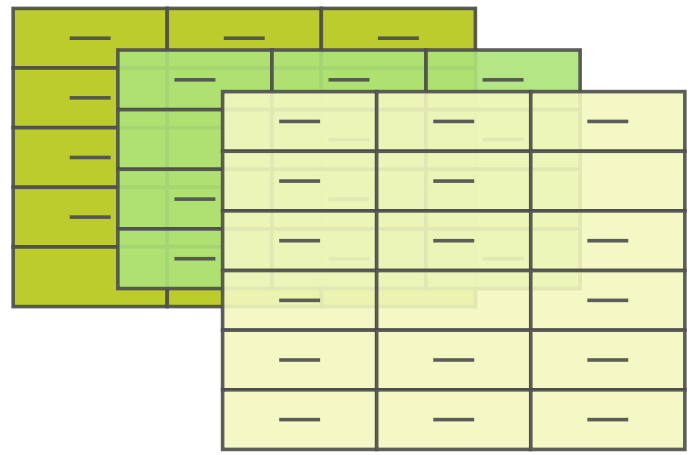
To scale Bayesian ER methods to millions of records without sacrificing accuracy and provide uncertainty of the ER task

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

Problem setup

Key assumptions:

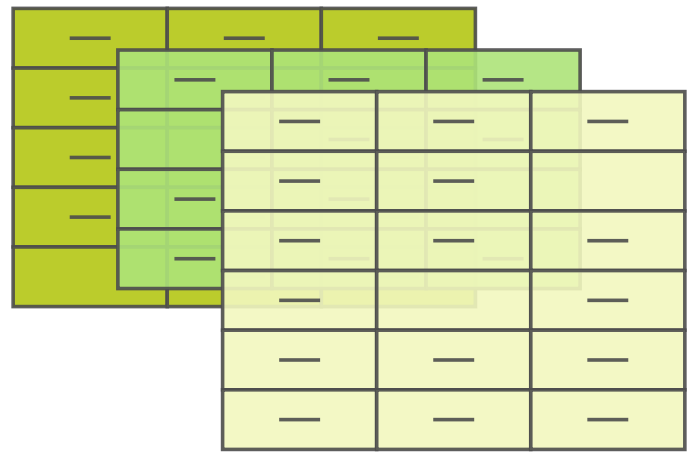
- multiple tables/sources
- duplicates within and across tables
- attributes are aligned
- attributes are discrete
- some missing values
- no ground truth (unsupervised)



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)



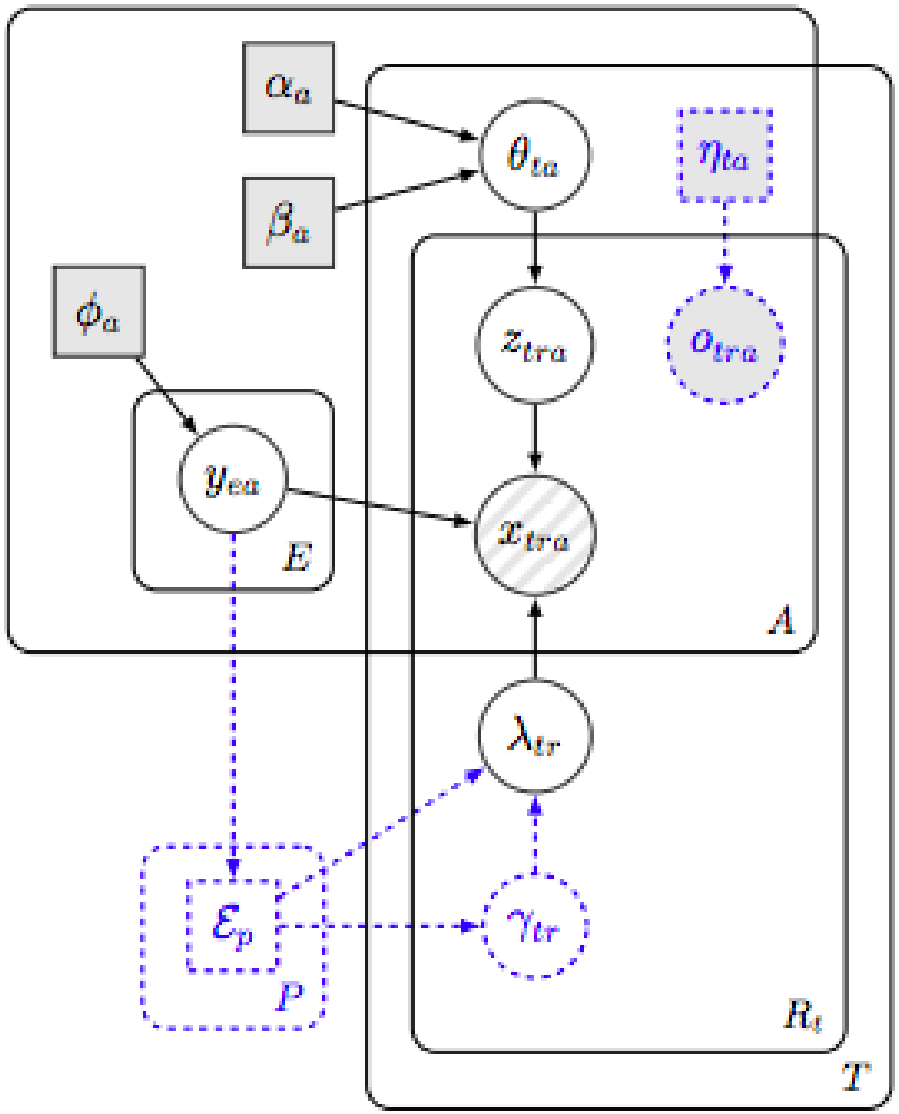
Output: approximate posterior distribution over the linkage structure

Our contribution

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, Kaplan, Rubinstein, Elazar, Steorts (2021)]

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.

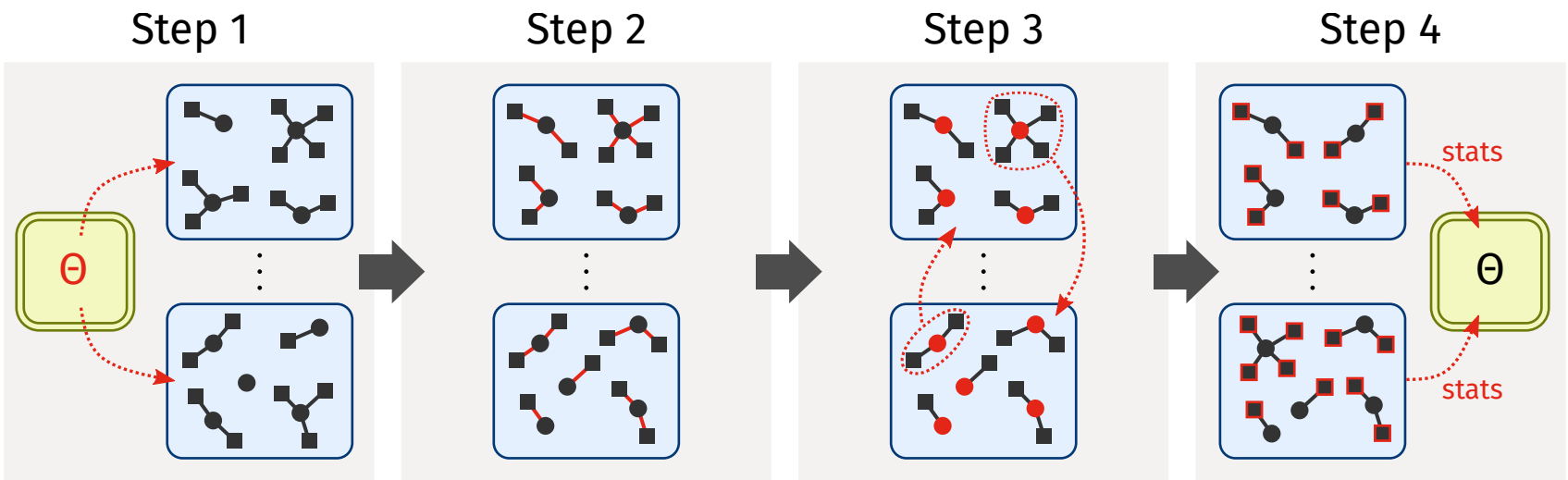
Distributed Markov chain Monte Carlo

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



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 Υ and Γ on the workers. Move the entities and records to their newly-assigned partitions.

Update Ζ, 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 dblink 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 dblink 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 1: 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

[Marchant+ (2021) JCGS]

Posterior Bias Plot

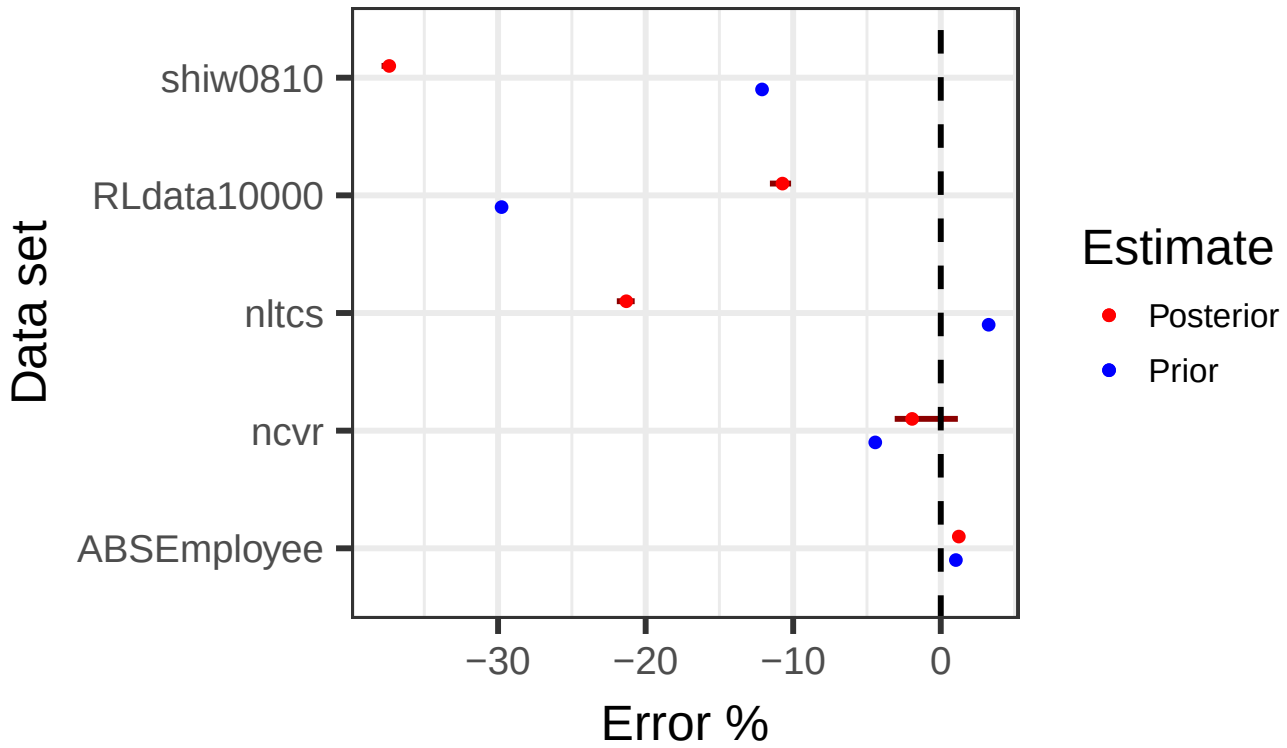
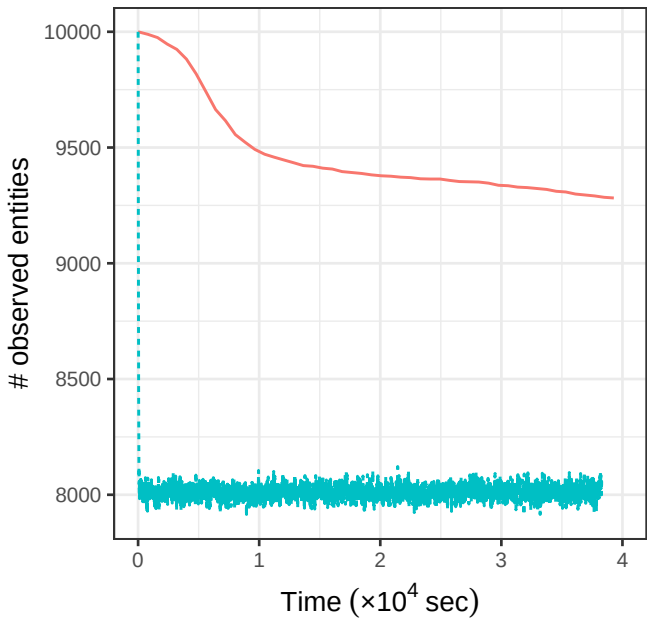


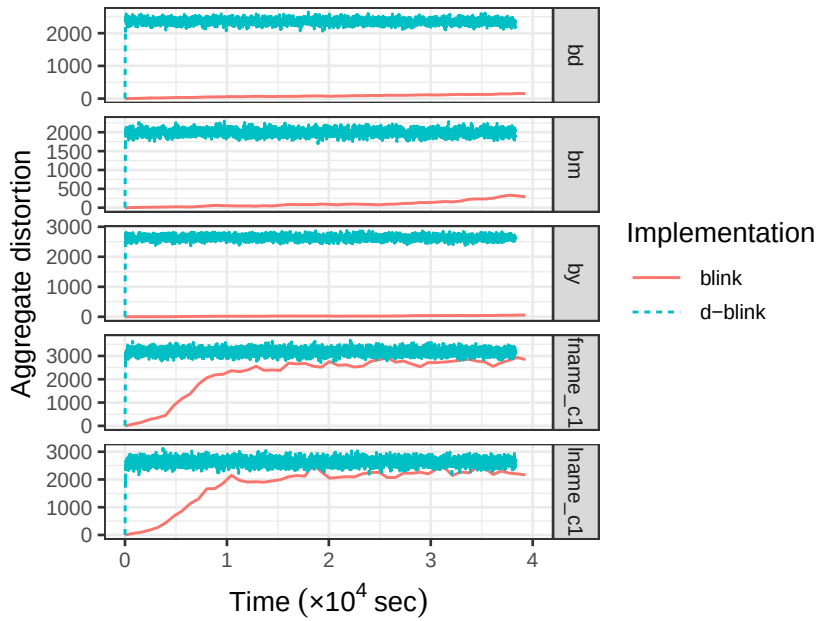
Figure 1: Error in the posterior and prior estimates for the number of observed entities for d-blink. The results show that the posterior estimate is very sharp and typically underestimates the true number, which is consistent with Steorts, Hall, Fienberg (2016).

[Marchant+ (2021) JCGS]

Convergence of d-blink versus blink

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

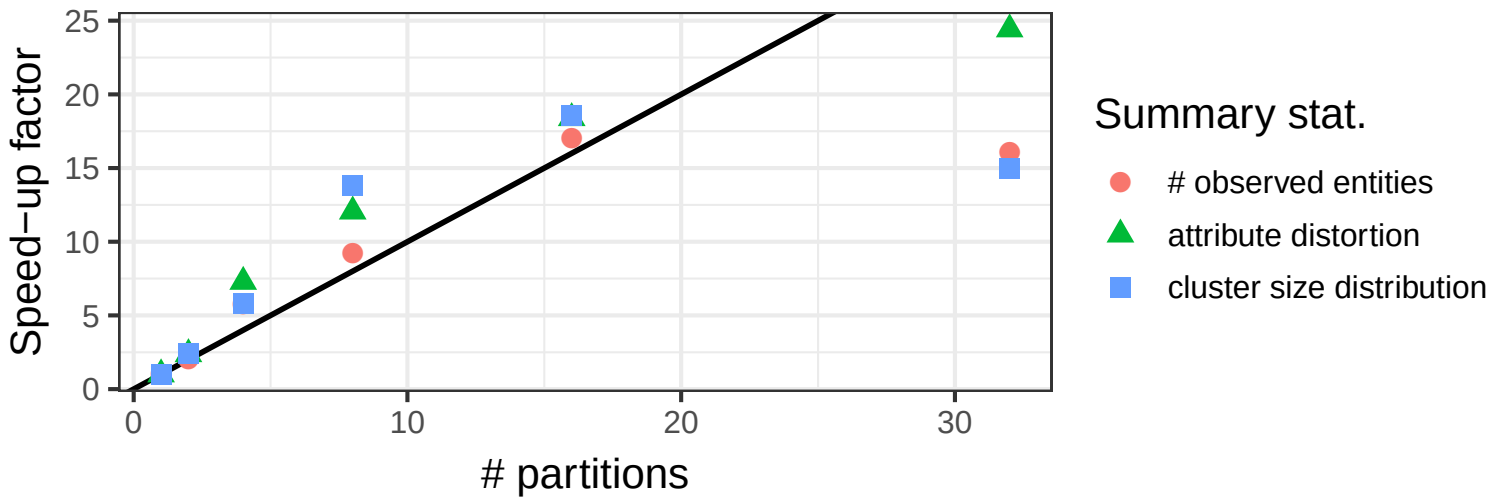




d-blink converges rapidly, however blink fails to reach the equilibrium distribution within 11 hours.

Does partitioning result in efficiency gains?

- Measure efficiency using ESS rate—the effective sample size generated per unit time
- Speed-up factor is the ESS rate relative to a baseline without partitioning
- Observe a near-linear speed-up for the `NLTCS` data set (tapering off beyond $\sim$ 20 partitions)



Case Study Applied to the 2010 Decennial Census

Table 2: Results for ER of 2010 Census and Numident data in Wyoming. Pairwise evaluation measures are computed using ground truth identifiers available for a subset of the records, where the unadjusted count was reported to be 563,626.

Pairwise measures			Posterior population size	
Precision	Recall	F1-score	Mean	Std. error
0.97	0.84	0.90	616,000	5,000

[Marchant+ (2021) JCGS]

Thank you!
Questions?

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

Binette and Steorts Review Article
Entity Resolution Software
Entity Resolution Tutorial