Will the Real Terry Tao Please Stand Up: A Bayesian Approach to Graphical Record Linkage

Rebecca C. Steorts

Department of Statistics
Carnegie Mellon University

joint with Rob Hall and Steve Fienberg

March 28, 2014

In conclusion

- Dealing with big data means merging large, noisy databases.

In conclusion

- Dealing with big data means merging large, noisy databases.
  - Medical data, official statistics, human rights violations, customer and transaction records, credit reports, ...

In conclusion

- Dealing with big data means merging large, noisy databases.
  - Medical data, official statistics, human rights violations, customer and transaction records, credit reports, ...
- Record linkage requires sophisticated graph structures.

In conclusion

- Dealing with big data means merging large, noisy databases.
  - Medical data, official statistics, human rights violations, customer and transaction records, credit reports, ...
- Record linkage requires sophisticated graph structures.
  - Use a bipartite graph for latent entities.

In conclusion

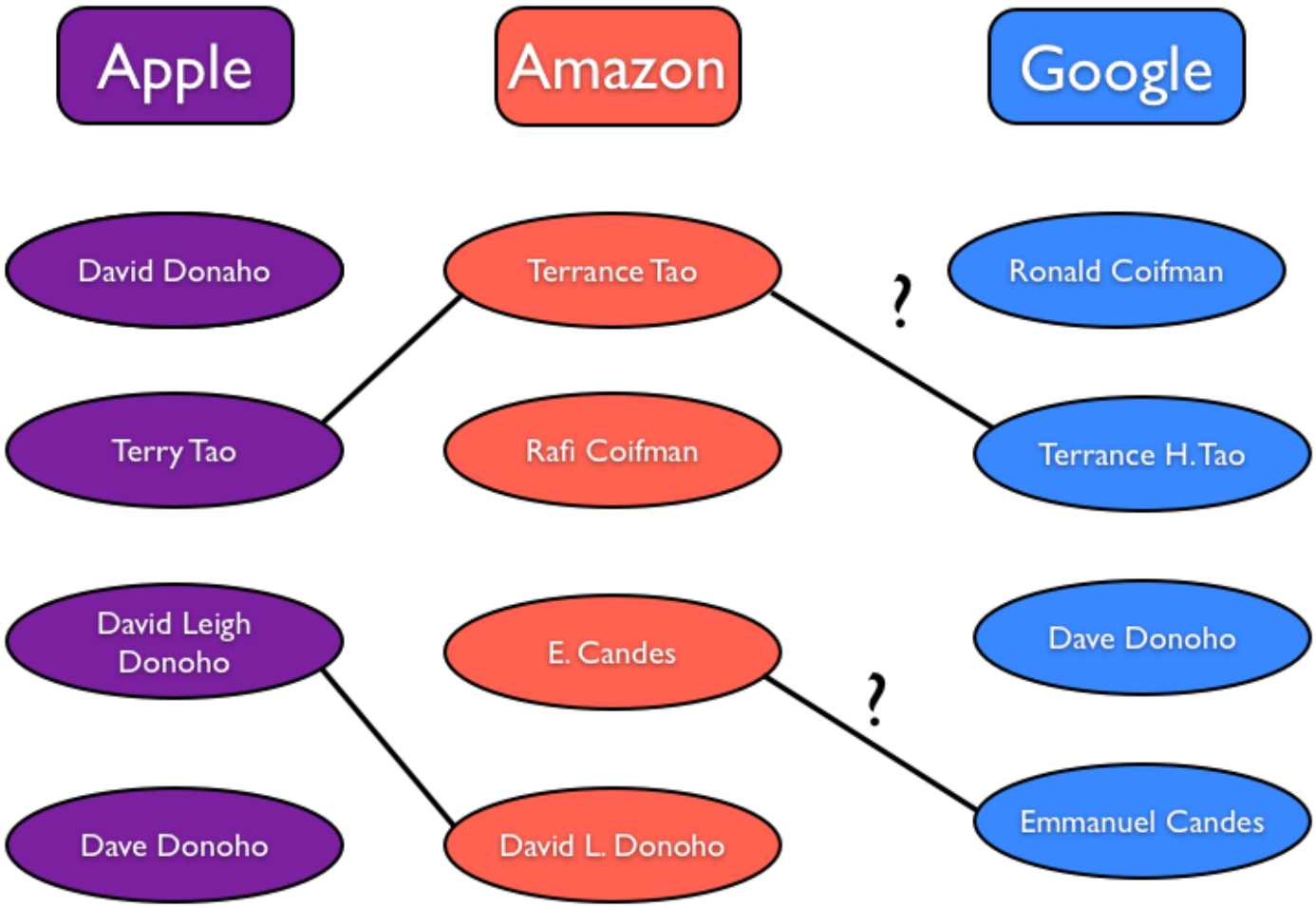
- Dealing with big data means merging large, noisy databases.
  - Medical data, official statistics, human rights violations, customer and transaction records, credit reports, ...
- Record linkage requires sophisticated graph structures.
  - Use a bipartite graph for latent entities.
  - Never link records to records.

In conclusion

- Dealing with big data means merging large, noisy databases.
  - Medical data, official statistics, human rights violations, customer and transaction records, credit reports, ...
- Record linkage requires sophisticated graph structures.
  - Use a bipartite graph for latent entities.
  - Never link records to records.
- We can estimate latent individuals across multiple high dimensional databases in linear time.

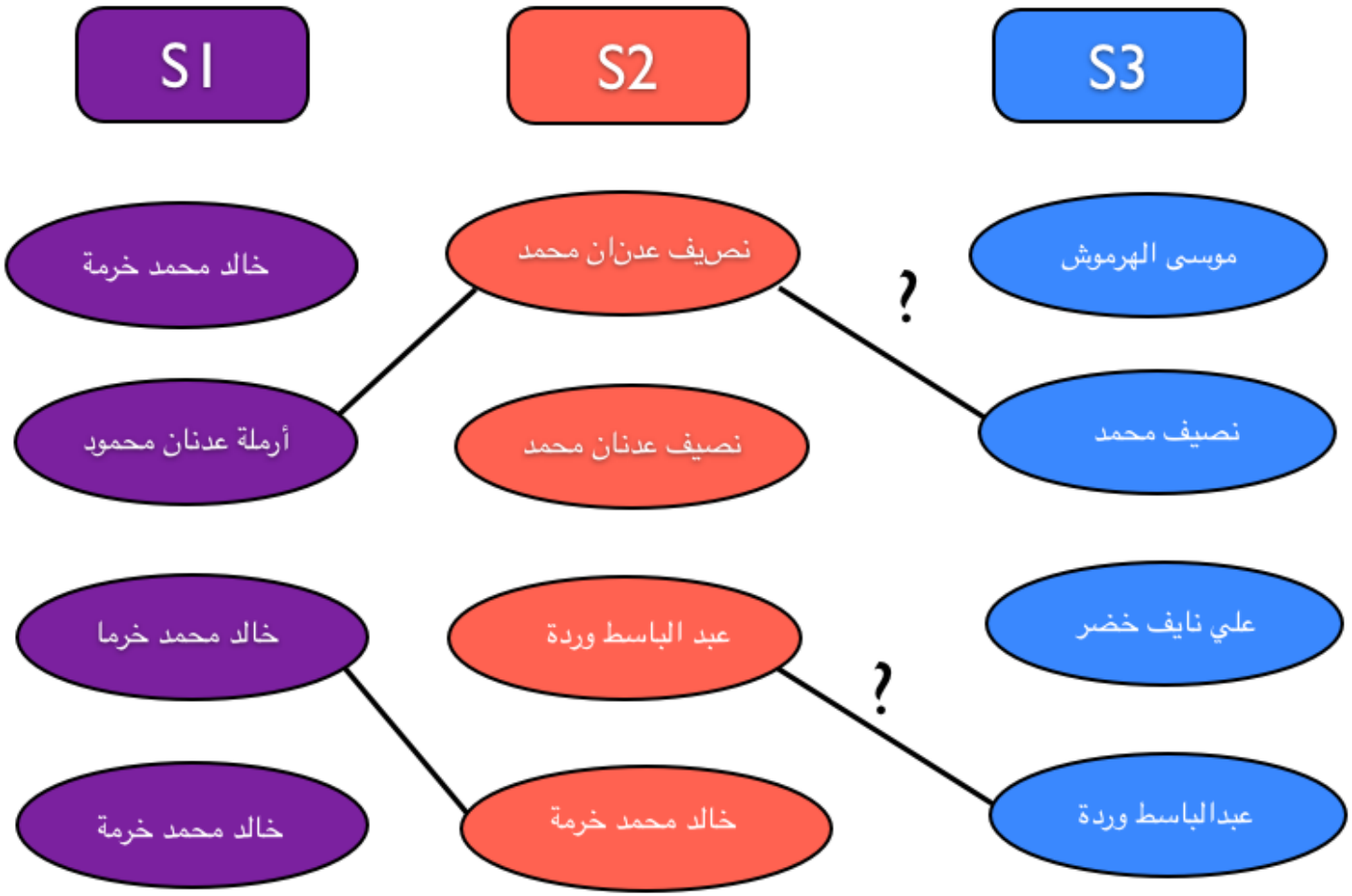
Record linkage joins multiple files without shared unique identifiers.

How do we find the "real" Terry Tao?



Link records to records. Learn edges using unsupervised modeling.

Syrian Civil War



My Frustrations, Let Me Show You Them

- Consider two files.
- Link records in file 1 to records in file 2.
- Treat deciding on a link as a hypothesis testing problem and estimate decision rules (Fellegi and Sunter 1969).
- Foundation for the vast majority of record linkage work.

My Frustrations, Let Me Show You Them

- Needs $O(N^2)$ record comparisons.
- No natural way of quantifying uncertainty.
- Awkward to extend to more than $k = 2$ files:
  - Hypothesis tests on triples (quadruples, k-tuples) of records with multiple alternative hypotheses.
  - Or just do pairwise linkage, and try to reconcile the different file pairs somehow.

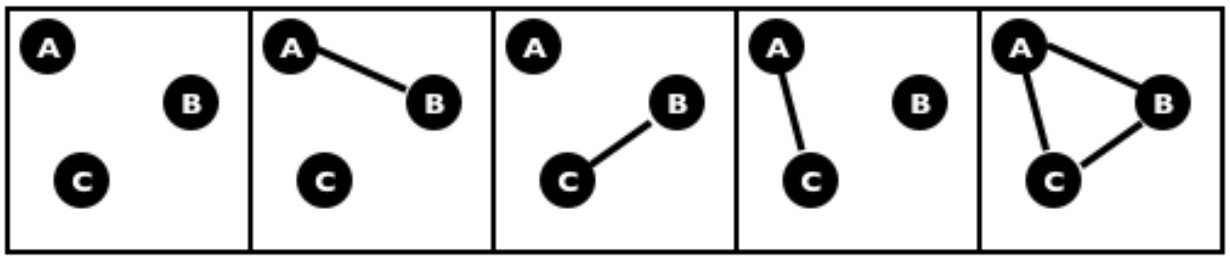
Transitivity

If A matches B and B matches C, then A ought to match C.

Transitivity

If A matches B and B matches C, then A ought to match C.

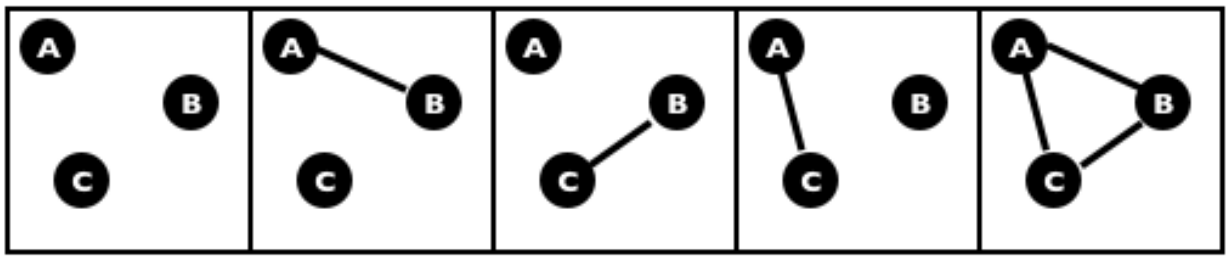
For three records, all the possible transitive relations.



Transitivity

If A matches B and B matches C, then A ought to match C.

For three records, all the possible transitive relations.



Pairwise hypothesis tests aren't necessarily transitive!

Transitivity

- Record linkage and transitivity resolved in Sadinle and Fienberg (2013) using k-tupled hypothesis testing.
- Computationally infeasible for moderate k.

- Bayesian literature: two files.
- Methods don't easily generalize to more than two files.
- Natural uncertainty quantification about links.

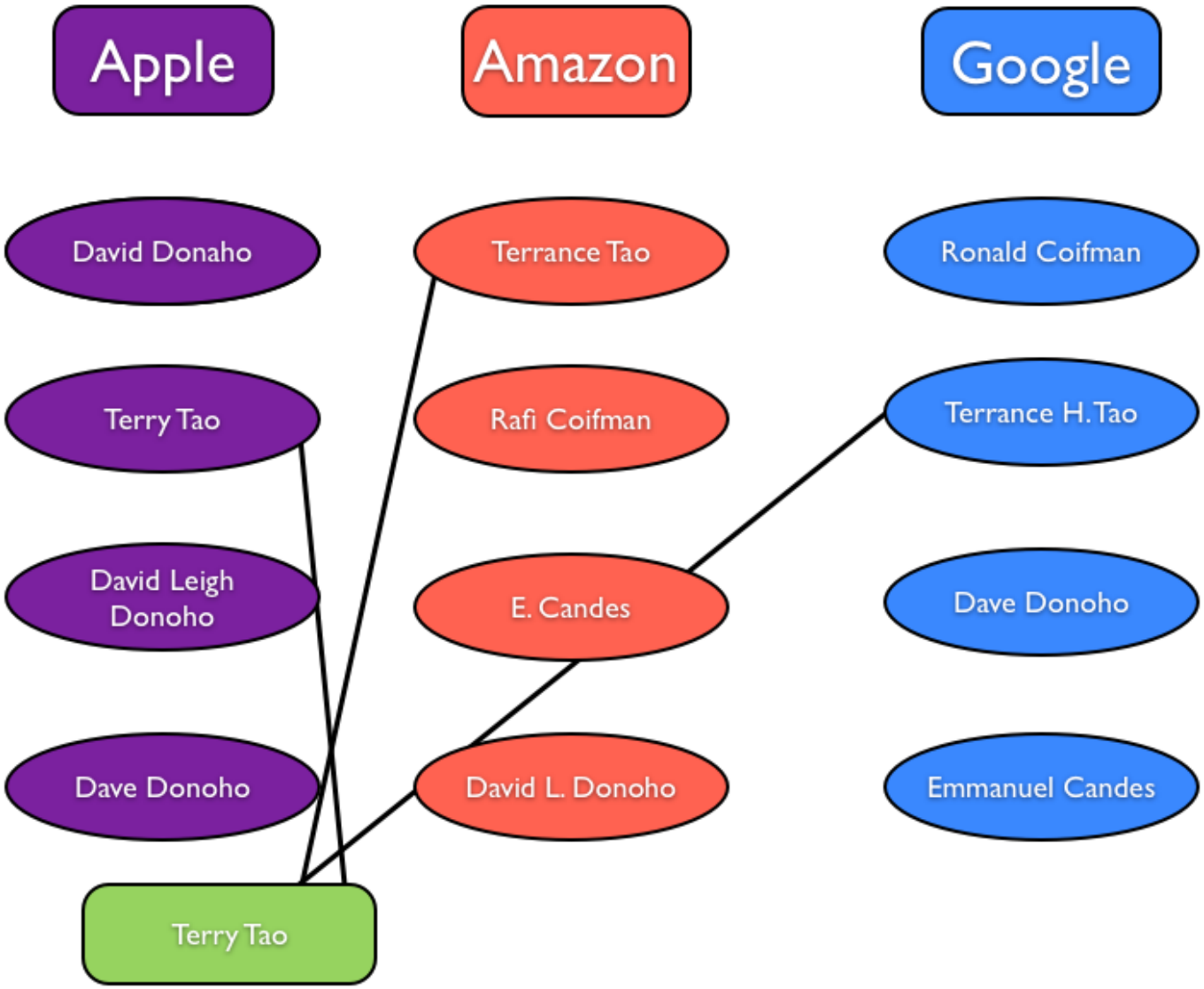
(Belin and Rubin 1995; Larsen and Rubin 2001; Tancredi and Liseo 2011; Gutman et al. 2013).

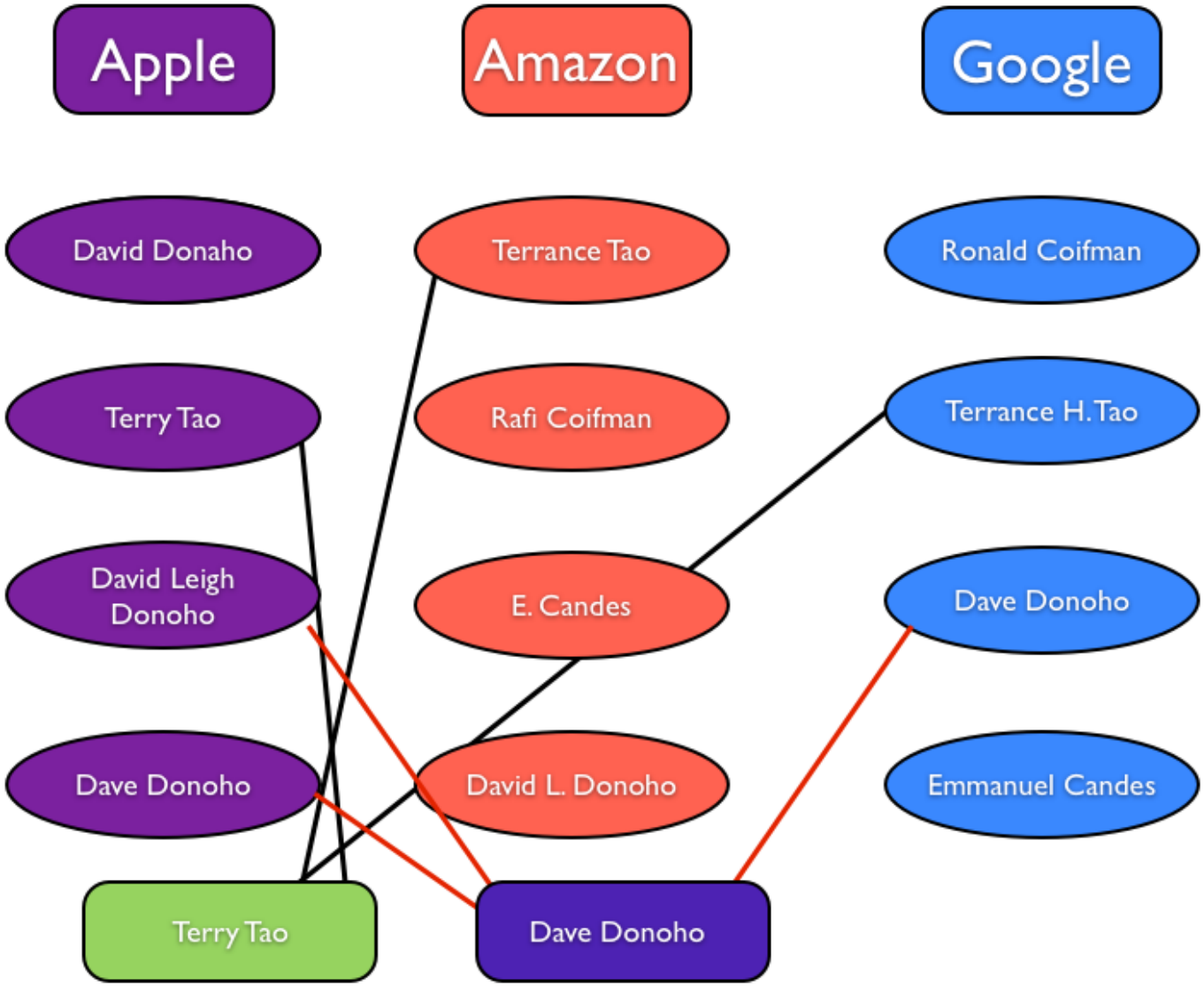
New approach: Link records to latent individuals.

Latent Individuals

- Latent individuals have attributes which can be compared to records.
- Records link only to latent individuals, never to other records.

  $\therefore$ bipartite graph.
- Transitivity is automatic.


Apple
Amazon
Google
David Donaho
Terrance Tao
Ronald Coifman
Terry Tao
Rafi Coifman
Terrance H. Tao
David Leigh Donoho
E. Candes
Dave Donoho
Dave Donoho
David L. Donoho
Emmanuel Candes
Terry Tao


Apple
Amazon
Google
David Donaho
Terrance Tao
Ronald Coifman
Terry Tao
Rafi Coifman
Terrance H. Tao
David Leigh Donoho
E. Candes
Dave Donoho
Dave Donoho
David L. Donoho
Emmanuel Candes
Terry Tao
Dave Donoho

- Our linkage graphs are novel data structures (record linkage and de-duplication).
- Can integrate our results into analytic procedures.
- Work in high dimensional parameter spaces.

Bayesian model: simultaneously does record linkage and de-duplication.

- Assume records are noisy.
- We have a novel data structure (links records to latent individuals).
- Hybrid MCMC.
- Algorithm is linear in records and MCMC iterations.
- Uncertainty quantification $\rightarrow$ leads to easy integration for analytical procedures.

[Steorts, Hall, and Fienberg (2014)]

Record linkage problems require a balance between computational speed and modeling.

Use Bayesian model with simplifying assumptions:

- Data is categorical.
- Lists are conditionally independent given the latent individuals.
- Number of fields within each list is the same.
- Assume that no field is *completely* missing for every record.

Notation

- $\mathbf{x}_{ij\ell}$ = data for ℓth field for jth record in file i.
- $\mathbf{y}_{j'\ell}$ = latent value for ℓth field of j'th record.
- $\mathbf{\Lambda}_{ij}$ = latent individual represented by jth record in file i.
- $\mathbf{z}_{ij\ell}$ whether or not the ℓth field for the jth record in file i is distorted.

$$\mathbf{x}_{ij\ell} \mid \mathbf{\Lambda}_{ij}, \mathbf{y}_{\mathbf{\Lambda}_{ij}\ell}, z_{ij\ell}, \boldsymbol{\theta}_\ell \overset{\text{ind}}{\sim} \begin{cases} \delta_{\mathbf{y}_{\mathbf{\Lambda}_{ij}\ell}} & \text{if } z_{ij\ell} = 0 \\ \text{Multinomial}(1, \boldsymbol{\theta}_\ell) & \text{if } z_{ij\ell} = 1 \end{cases}$$

$$z_{ij\ell} \mid \beta_\ell \overset{\text{ind}}{\sim} \text{Bernoulli}(\beta_\ell)$$

$$\mathbf{y}_{j'\ell} \mid \boldsymbol{\theta}_\ell \overset{\text{ind}}{\sim} \text{Multinomial}(1, \boldsymbol{\theta}_\ell)$$

$$\boldsymbol{\theta}_\ell \overset{\text{ind}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_\ell)$$

$$\beta_\ell \overset{\text{ind}}{\sim} \text{Beta}(a_\ell, b_\ell)$$

$$\pi(\mathbf{\Lambda}) \propto 1.$$

$\mathbf{\Lambda}$ represents the linkage structure:

- Records (i_1, j_1) and (i_2, j_2) refer to same individual $\Leftrightarrow$ $\mathbf{\Lambda}_{i_1 j_1} = \mathbf{\Lambda}_{i_2 j_2}$.
- Indicates links across files when $i_1 \neq i_2$ ("record linkage").
- Indicates links within files when $i_1 = i_2$ ("de-duplication").

$\mathbf{\Lambda}$ represents the linkage structure:

- Records (i_1, j_1) and (i_2, j_2) refer to same individual $\Leftrightarrow$ $\mathbf{\Lambda}_{i_1 j_1} = \mathbf{\Lambda}_{i_2 j_2}$.
- Indicates links across files when $i_1 \neq i_2$ ("record linkage").
- Indicates links within files when $i_1 = i_2$ ("de-duplication").

$\mathbf{\Lambda}$ helps us estimate:

- Total number of unique individuals surveyed.
- Which individuals appear in multiple files.
- Capture probabilities as functions of attributes.

$\mathbf{\Lambda}$ represents the linkage structure:

- Records (i_1, j_1) and (i_2, j_2) refer to same individual $\Leftrightarrow$ $\mathbf{\Lambda}_{i_1 j_1} = \mathbf{\Lambda}_{i_2 j_2}$.
- Indicates links across files when $i_1 \neq i_2$ ("record linkage").
- Indicates links within files when $i_1 = i_2$ ("de-duplication").

$\mathbf{\Lambda}$ helps us estimate:

- Total number of unique individuals surveyed.
- Which individuals appear in multiple files.
- Capture probabilities as functions of attributes.

$\mathbf{\Lambda}$ is a high-dimensional parameter:

$$\dim(\mathbf{\Lambda}) = N_{\text{max}}.$$

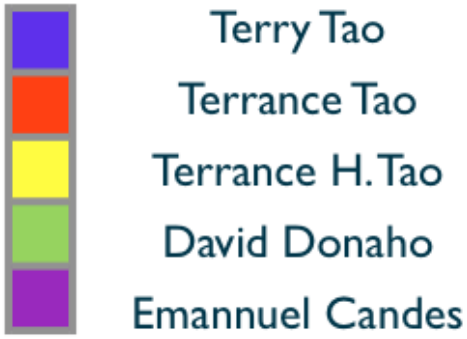
Record linkage problems demand fast algorithms and computational speed ups.

We cluster records to latent individuals.

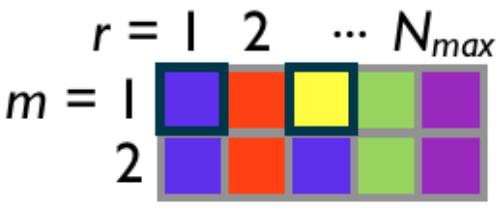
- Perform hybrid Markov chain Monte Carlo (MCMC).
- Multiple Metropolis iterations within Gibbs sampling.
  - Uniformly draw pairs of records within each Metropolis step.
  - Then either Split or Merge them.
  - Uses the method of Jain and Neal (2004).

Split Merge Procedure

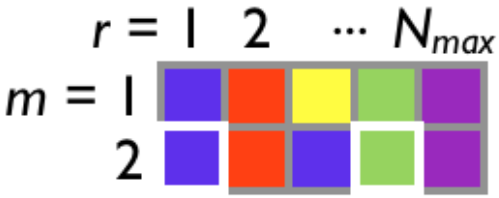
We initialize each record to a latent individual (indicated by color).



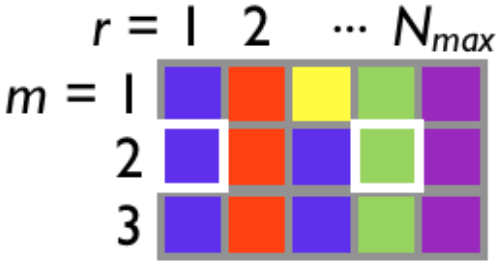
At iteration 1, we propose merging two latent individuals.



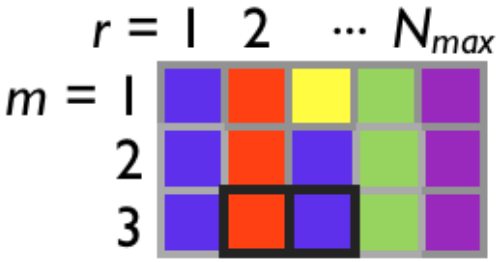
After iteration 1, we have merged two latent individuals.



At iteration two, we propose merging two latent individuals.



After iteration 2, we have rejected the merge.



At iteration 3, we propose merging two latent individuals.

After iteration 3, we merge a blue and a red, resulting in two blue individuals.

At iteration 4, we propose a split.

After iteration 4, we have split two blue individuals into blue and yellow latent individuals.

1. Start out with records initialized to their index $(1, \ldots, N_{\text{max}})$. For $m = 1, 2, \ldots, S_G$ (Gibbs):
2. At each Gibbs iteration, split or merge latent individuals. Repeat.
3. Resample parameters; repeat Gibbs.

Blocking

Blocking is a speed up in record linkage, where reliable fields are treated as "fixed" or not distorted.

- We only compare records within blocks.
- Avoids all to all comparisons.
- May be unrealistic for some datasets.

Oh, Convergence!

- Geometric ergodicity follows from Jain and Neal (2004).
  - Recall that a chain is geometrically ergodic if $\exists B$ and $r > 0$

$$\|P^m(x, \cdot) - \pi(\cdot)\|_1 \leq B(x)e^{-mr} \ \forall x \in \mathcal{X}.$$

- Algorithm is linear in the number of records and MCMC iterations.
- *Bound* on convergence rate much harder: how does r scale with $N_{\max}$? — current work.

Time Complexity

- Algorithm is linear in both $N_{\max}$ and MCMC iterations.
- Let $M = \frac{1}{p}\sum_{\ell=1}^{p} M_\ell$ as the average number of possible values per field ($M \geq 1$).

- For S_G iterations of the Gibbs sampler, the algorithm is order $O(pMN_{\max}S_GS_M)$.
- If (p and M) $\ll N_{\max}$, the runtime is $O(N_{\max}S_GS_M)$.

For details see Steorts, Hall, and Fienberg (2014).

How does our matching perform overall?
What about errors?

Posterior Matching Probabilities

Records (i_1, j_1) and (i_2, j_2) *match* if they point to the same latent individual, i.e., if $\mathbf{\Lambda}_{i_1 j_1} = \mathbf{\Lambda}_{i_2 j_2}$. This implies that

$$P(\mathbf{\Lambda}_{i_1 j_1} = \mathbf{\Lambda}_{i_2 j_2} | \mathbf{X}) = \frac{1}{S_G} \sum_{h=1}^{S_G} I(\mathbf{\Lambda}_{i_1 j_1}^{(h)} = \mathbf{\Lambda}_{i_1 j_2}^{(h)}).$$

We often need to report a point estimate of the entire linkage structure.

- A set of records $\mathcal{A}$ is a *maximal matching* set (MMS) if:
  - every record in the set has the same value of $\mathbf{\Lambda}_{ij}$ and
  - no record outside the set has that value of $\mathbf{\Lambda}_{ij}$.
- Define $\mathbf{f}(\mathcal{A}, \mathbf{\Lambda})$ to be 1 if $\mathcal{A}$ is an MMS in $\mathbf{\Lambda}$ and 0 otherwise:

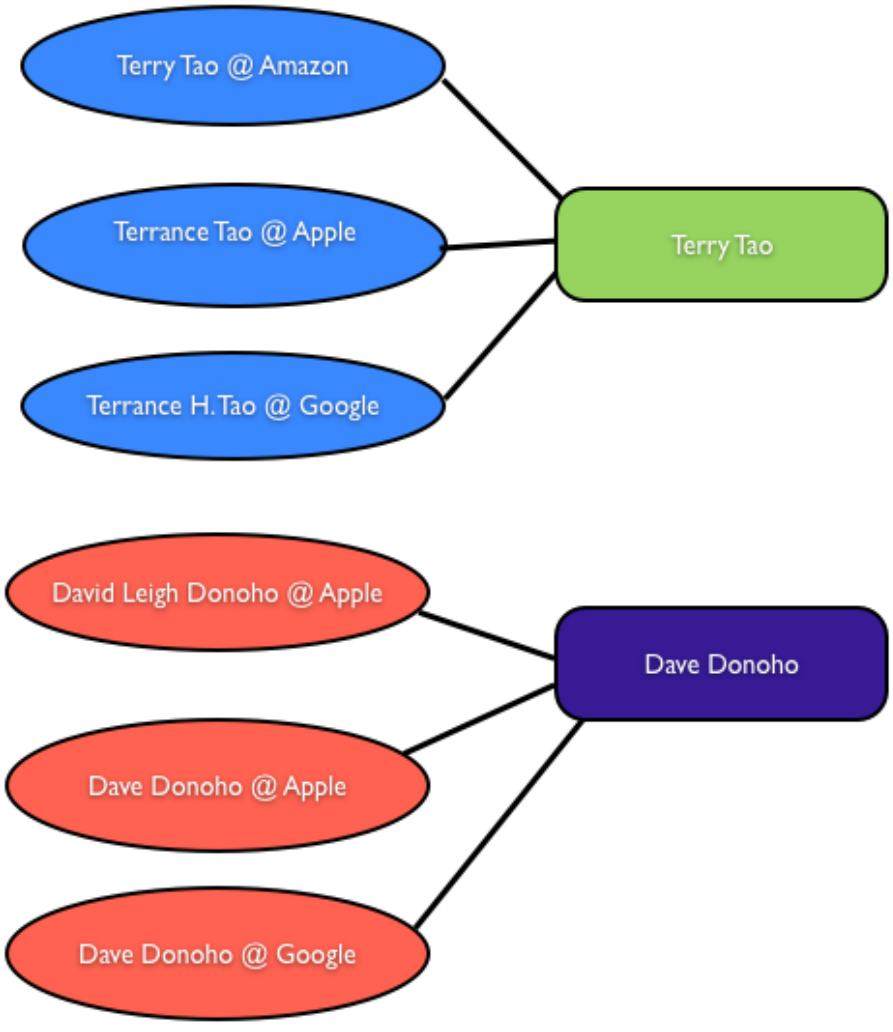
- A set of records $\mathcal{A}$ is a *maximal matching* set (MMS) if:
  - every record in the set has the same value of $\mathbf{\Lambda}_{ij}$ and
  - no record outside the set has that value of $\mathbf{\Lambda}_{ij}$.
- Define $\mathbf{f}(\mathcal{A}, \mathbf{\Lambda})$ to be 1 if $\mathcal{A}$ is an MMS in $\mathbf{\Lambda}$ and 0 otherwise:

$$\mathbf{f}(\mathcal{A}, \mathbf{\Lambda}) = \sum_{j'} \left(\prod_{(i,j) \in \mathcal{A}} I(\mathbf{\Lambda}_{ij} = j') \prod_{(i,j) \notin \mathcal{A}} I(\mathbf{\Lambda}_{ij} \neq j') \right).$$

- Records are in the same MMS $\iff$ they match the same latent individual.
  - Which individual is irrelevant.

- Records are in the same MMS $\iff$ they match the same latent individual.
  - Which individual is irrelevant.
- For a set of records $\mathcal{A}$, the posterior probability it is an MMS:

$$P(\mathbf{f}(\mathcal{A}, \mathbf{\Lambda}) = 1 \mid \mathbf{X}) = \frac{1}{S_G} \sum_{h=1}^{S_G} \mathbf{f}(\mathcal{A}, \mathbf{\Lambda}^{(h)}).$$



Illustrating two MMSs.

- MMSs: preserve transitivity.
- *Most probable MMS* (MPMMSs) $\mathcal{M}_{ij}$: set containing record (i, j) with the highest posterior probability of being an MMS, i.e.,

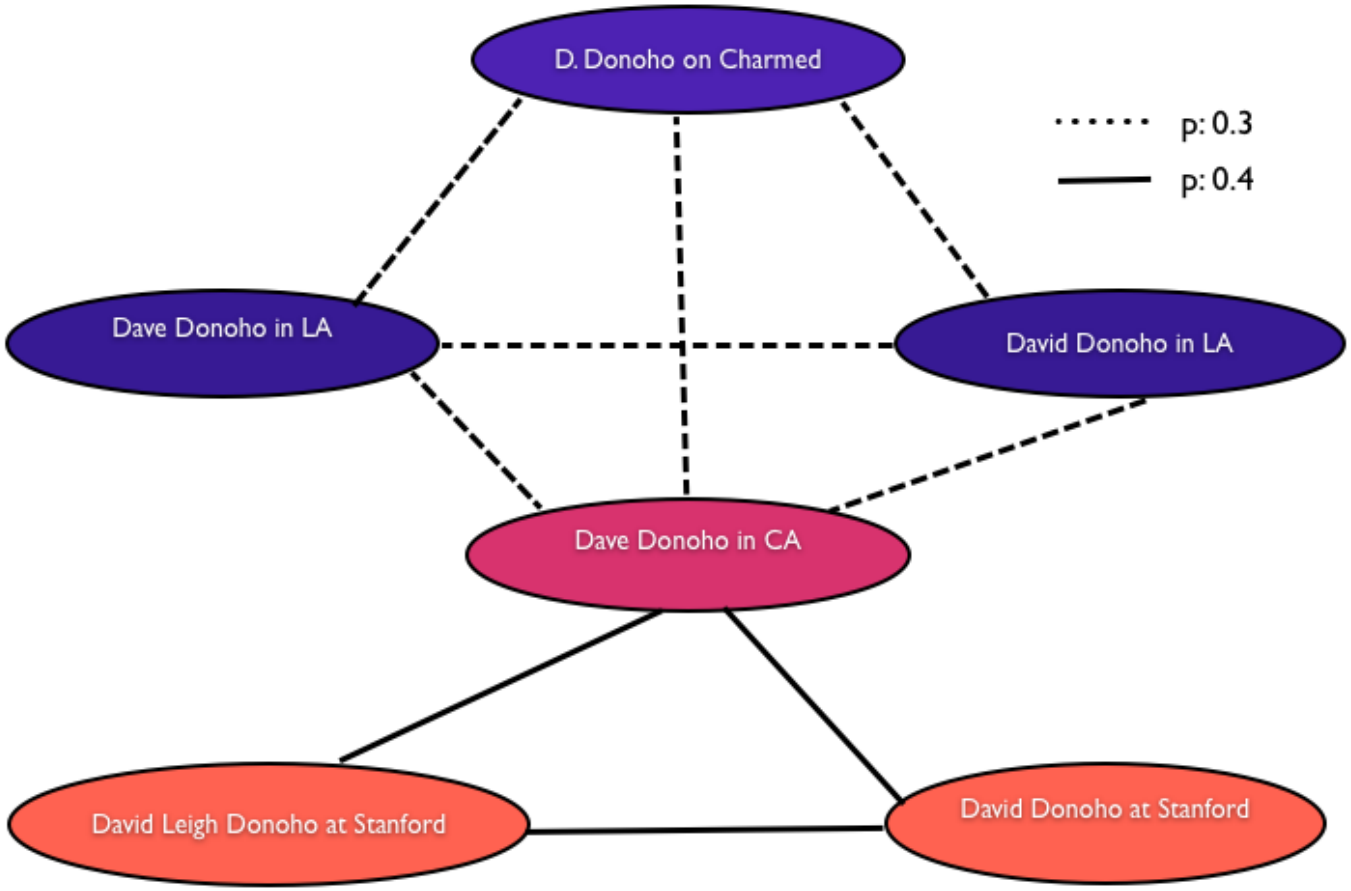
- MMSs: preserve transitivity.
- *Most probable MMS* (MPMMSs) $\mathcal{M}_{ij}$: set containing record (i, j) with the highest posterior probability of being an MMS, i.e.,

$$\mathcal{M}_{ij} := \underset{\mathcal{A}:(i,j)\in\mathcal{A}}{arg\ max}\ P(\mathbf{f}(\mathcal{A}, \mathbf{\Lambda}) = 1 \mid \mathbf{X}).$$

- MMSs: preserve transitivity.
- *Most probable MMS* (MPMMSs) $\mathcal{M}_{ij}$: set containing record (i, j) with the highest posterior probability of being an MMS, i.e.,

$$\mathcal{M}_{ij} := \underset{\mathcal{A}:(i,j)\in\mathcal{A}}{arg\ max}\ P(\mathbf{f}(\mathcal{A}, \mathbf{\Lambda}) = 1 \mid \mathbf{X}).$$

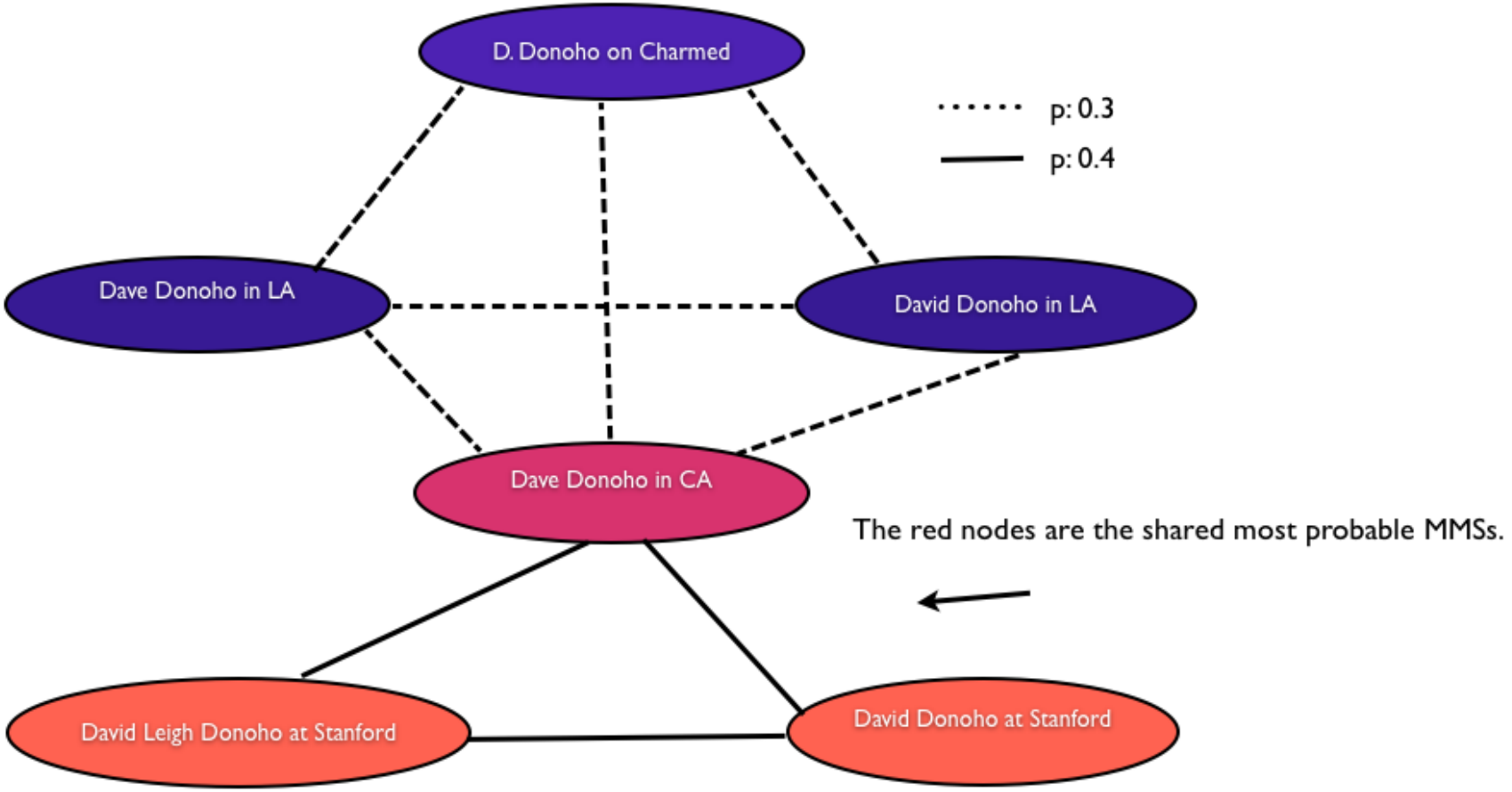
- Issue: possible for different records' most probable MMSs to contradict each other.



MPMMSs can violate transitivity.

Resolving Transitivity

- *Shared MPMMS*: set that is the MPMMS for each of its members.
- Estimate the overall linkage structure by linking records $\iff$ they are in the same shared MPMMS.


D. Donoho on Charmed
p: 0.3
p: 0.4
Dave Donoho in LA
David Donoho in LA
Dave Donoho in CA
The red nodes are the shared most probable MMSs.
David Leigh Donoho at Stanford
David Donoho at Stanford

- National Long Term Care Survey (NLTCS): survey about health/functional status of Americans 65 + .
- Approximately 20,000 individuals/wave.
- Test the record matching on three of the six waves.
  - Due to survey design, three waves are not useable for linkage.
  - Over the survey, patients either die or drop out. Replace at the next wave.
- Data used here: DOB, sex, state, and office location.
- Block on year of birth and sex.
  - Similar results when we do not block.
- Have unique IDs.

Posterior Matching Probabilities for NLTCS

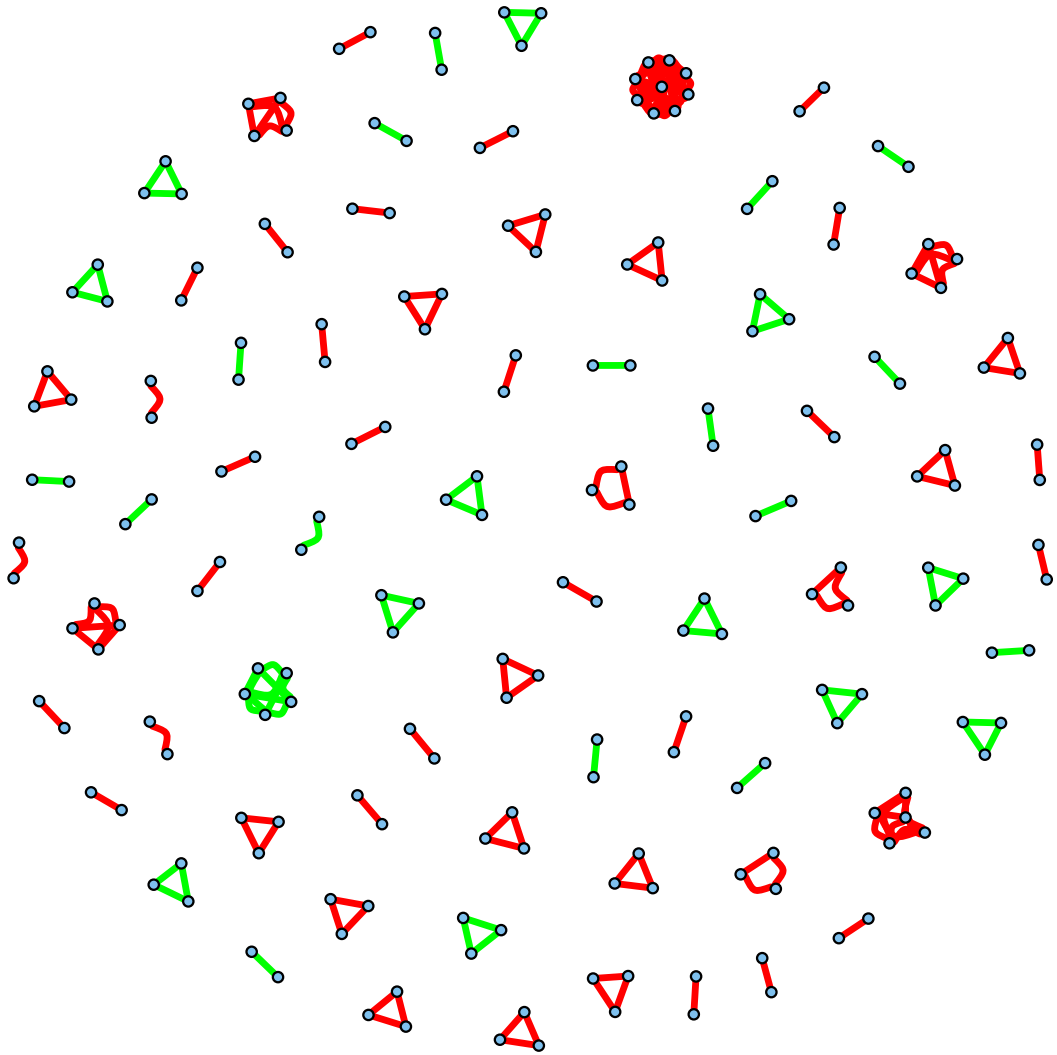
- Find posterior matching probabilities using $\mathbf{\Lambda}$.
- Suppose we're interested in patient 10084 in wave 1:

sets of records	1.10084	3.5583; 1.10084	3.5583; 1.10084; 2.6131
posterior probability	0.001	0.004	0.995

The set of three records 3.5583; 1.10084; 2.6131 all agree:

- Male with DOB: 07/XX/XX.
- Visited office 25.
- From Illinois.

and ground truth says they are the same individual.



Every node is a record and each edge is a link between records.

Groups are transitive.

Split and MErge REcord linkage and De-duplication

- Two variants of algorithm: SMERED and SMERE.
- Assess both using shared most probable MMS's.

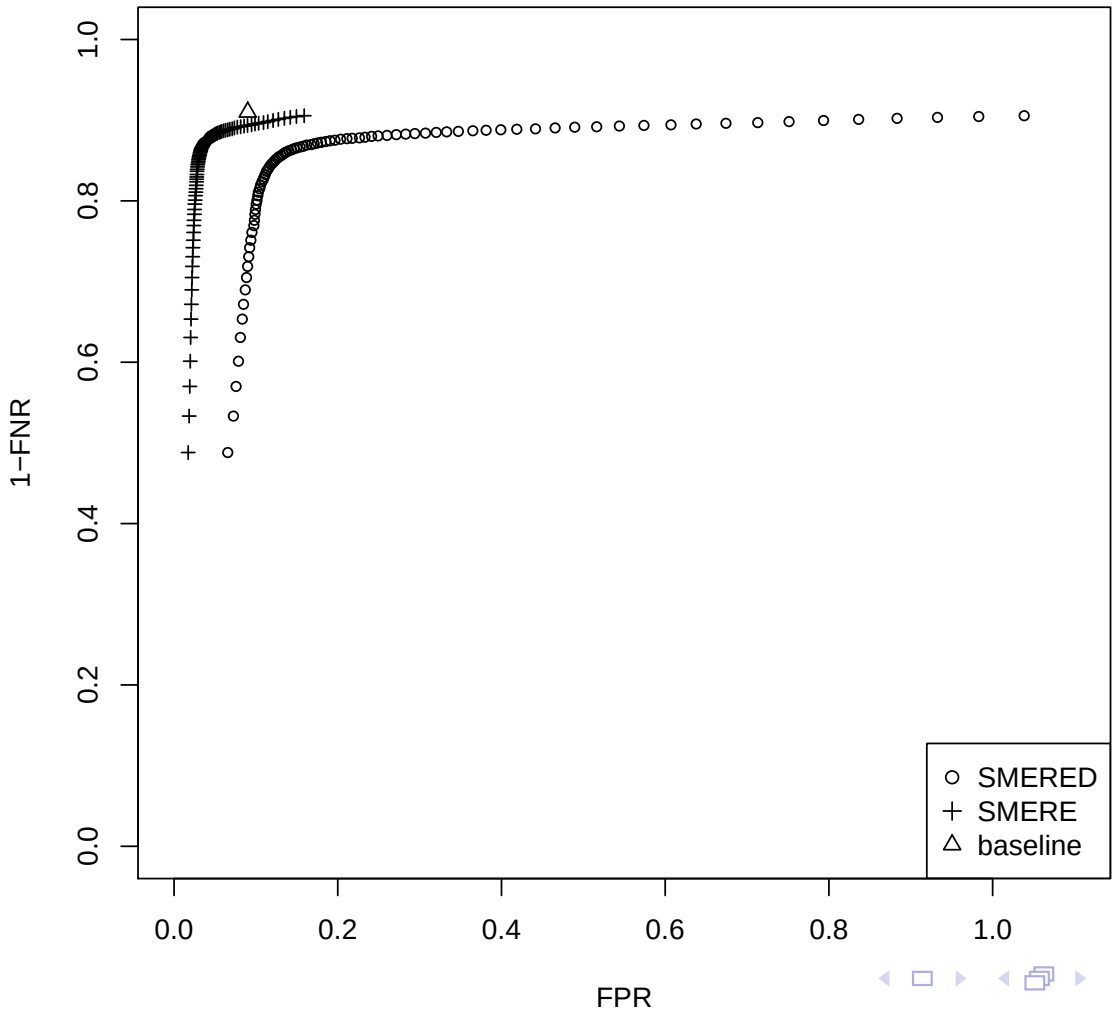
Note that:

$$\text{Splitting or FNR} = \frac{\#\text{ false negatives}}{\#\text{ false negatives} + \#\text{true positives}} \text{ and}$$

$$\text{Lumping or FPR} = \frac{\#\text{ false positives}}{\#\text{ false negatives} + \#\text{true positives}}.$$

Splitting: prevalence of false negative matches, and lumping: prevalence of false positive matches.

Split and MErge REcord linkage and De-duplication

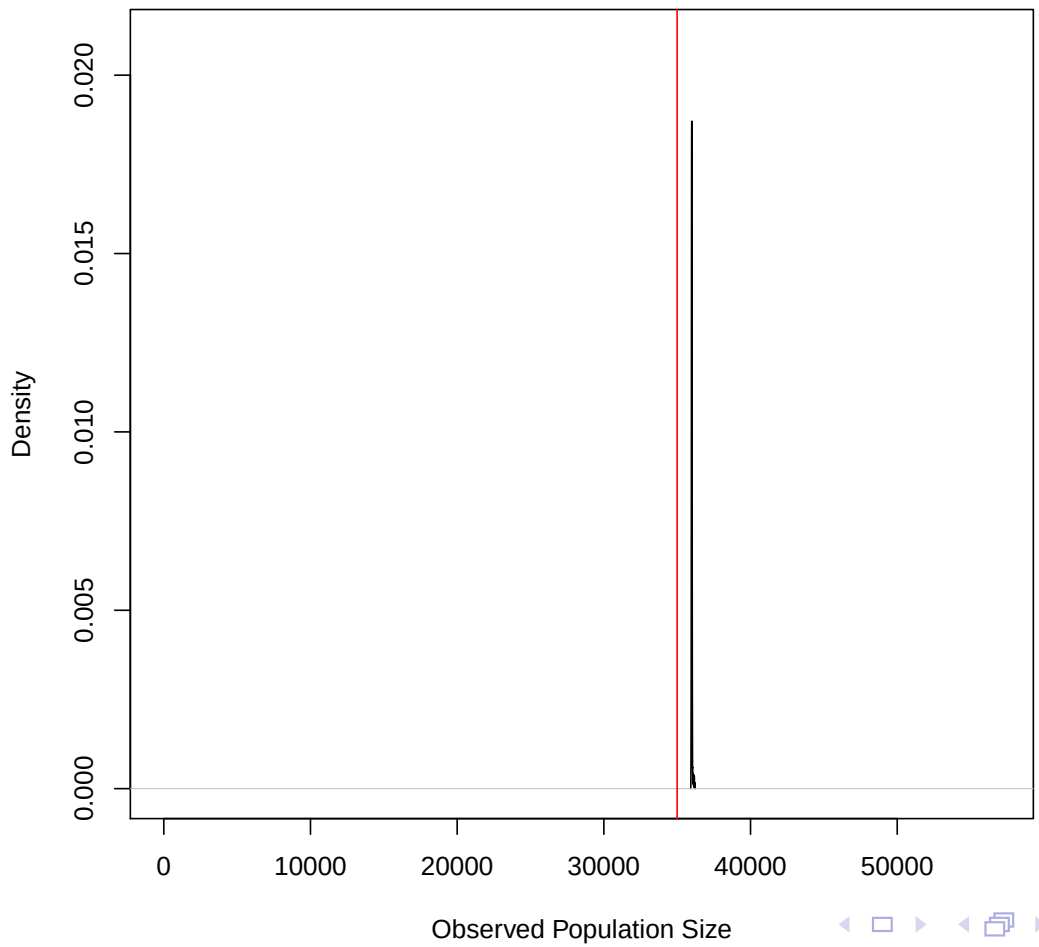


Why Record Linkage and De-duplication?

SMERED does better when there are duplicates within lists.
NLTCS was carefully designed to have no duplicates within lists.

Inferring the Number of Observed Individuals N

- $E(N|\mathbf{X}) = 35,992$ with a posterior standard error of 19.08.
- The true population is 34,945 $\rightarrow$ undermatching.



- Clusters records to latent individuals using an EB approach.
- Incorporates both categorical and string data.

$$\Pr(\text{observe string } w' \mid \text{true string } w) \propto \Pr_{\text{empirical}}(w')\, e^{-c\, d(w',w)}.$$

- The string metric d is general and decided by the user.
- All the full conditionals are in closed form.

Steorts (2014)

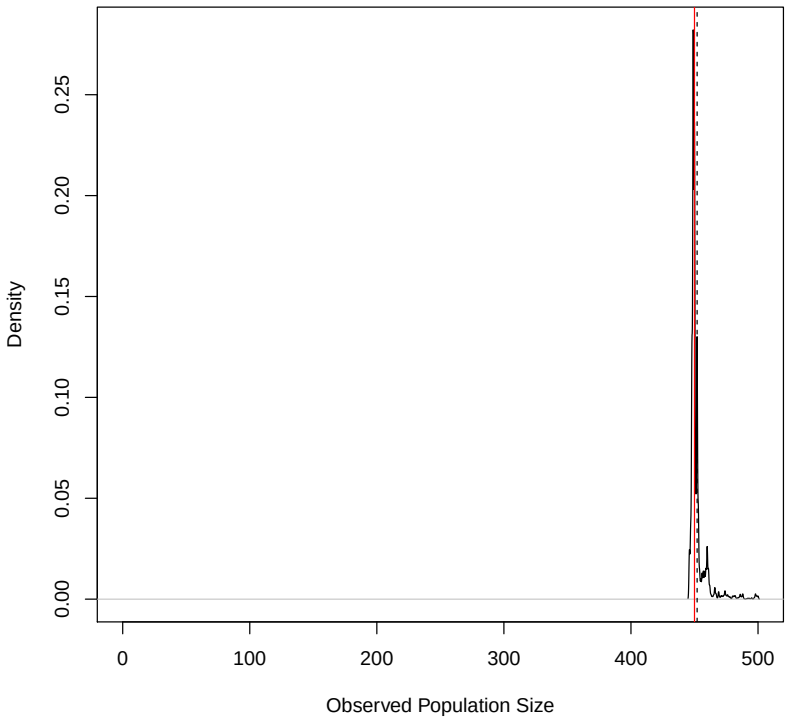


Figure : Posterior density for N in simulation study. The FNR and FPR: 0.04 and 0.02.

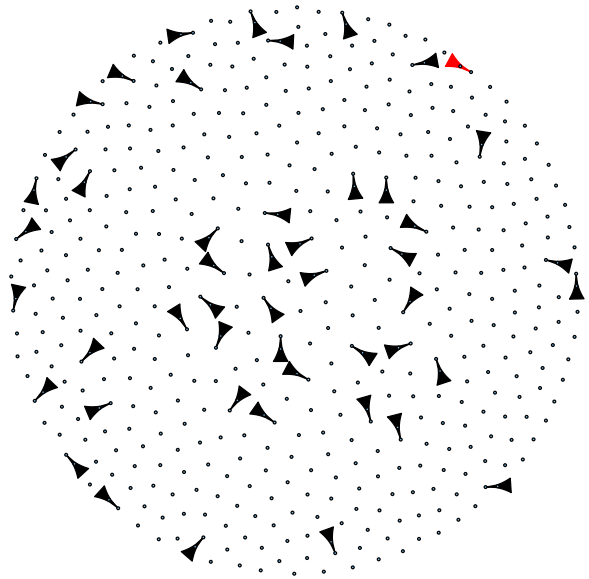


Figure : Shared MPMMS graphical representation of simulation study. Only makes one false positive set.

Extensions:

1. Empirical applications to human rights data (Steorts, 2014).

Extensions:

1. Empirical applications to human rights data (Steorts, 2014).
2. The latent individuals are not exchangeable.
   - Classical models on the latents is not appropriate. (Ciollaro and Steorts, 2014).

Extensions:

1. Empirical applications to human rights data (Steorts, 2014).
2. The latent individuals are not exchangeable.
   - Classical models on the latents is not appropriate. (Ciollaro and Steorts, 2014).

Extensions:

1. Empirical applications to human rights data (Steorts, 2014).
2. The latent individuals are not exchangeable.
   - Classical models on the latents is not appropriate. (Ciollaro and Steorts, 2014).

Extensions:

1. Empirical applications to human rights data (Steorts, 2014).
2. The latent individuals are not exchangeable.
   - Classical models on the latents is not appropriate. (Ciollaro and Steorts, 2014).
3. Theoretical Limits on record linkage.
   - Record are noisy and latents appear a finite number of times.
   - Can we use information theory, combinatorics to put upper or lower limits on how much we can learn on the latents?
   - Is there something like sparsity for record linkage?

Extensions:

1. Empirical applications to human rights data (Steorts, 2014).
2. The latent individuals are not exchangeable.
   - Classical models on the latents is not appropriate. (Ciollaro and Steorts, 2014).
3. Theoretical Limits on record linkage.
   - Record are noisy and latents appear a finite number of times.
   - Can we use information theory, combinatorics to put upper or lower limits on how much we can learn on the latents?
   - Is there something like sparsity for record linkage?
4. Privacy-preserving record linkage: goal is to limit how much information is revealed about individual-level data by linkage.

Extensions:

1. Empirical applications to human rights data (Steorts, 2014).
2. The latent individuals are not exchangeable.
   - Classical models on the latents is not appropriate. (Ciollaro and Steorts, 2014).
3. Theoretical Limits on record linkage.
   - Record are noisy and latents appear a finite number of times.
   - Can we use information theory, combinatorics to put upper or lower limits on how much we can learn on the latents?
   - Is there something like sparsity for record linkage?
4. Privacy-preserving record linkage: goal is to limit how much information is revealed about individual-level data by linkage.
   - How does the record linkage algorithm degrade as we have less data available (and we inject noise into the process)?
   - What is the utility guaranteed under our methods?

Theme of My Work

- Recover a high dimensional object, such as a latent individual or measurement for a small domain from degraded data.
- How to use this data and recover the underlying structure?

Thanks to the NSF Census Research Group at CMU and funding from the NSF.

Questions?

beka@cmu.edu

Thomas R Belin and Donald B Rubin. A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90(430):694–707, 1995.

I. Fellegi and A. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

R. Gutman, C. Afendulis, and A. Zaslavsky. A bayesian procedure for file linking to analyze end- of-life medical costs. *Journal of the American Statistical Association*, 108(501):34–47, 2013.

S. Jain and R. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13:158–182, 2004.

Michael D Larsen and Donald B Rubin. Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96(453):32–41, 2001.

M. Sadinle and S.E. Fienberg. A generalized Fellegi-Sunter framework for multiple record linkage with application to homicide record-systems. *Journal of the American Statistical Association*, 108(502):385–397, 2013.

R. C. Steorts, Rob Hall, and S.E. Fienberg. A Bayesian approach to graphical record linkage and de-duplication. *Submitted*, 2013.

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

New Work: Empirical Bayes Model

- Define $\alpha_\ell(w) =$ relative frequency of w in data for field ℓ.

New Work: Empirical Bayes Model

- Define $\alpha_\ell(w) =$ relative frequency of w in data for field ℓ.
- G_ℓ: empirical distribution for field ℓ.

New Work: Empirical Bayes Model

- Define $\alpha_\ell(w) =$ relative frequency of w in data for field ℓ.
- G_ℓ: empirical distribution for field ℓ.
- $W \sim F_\ell(w_0)$:

$$P(W = w) \propto \alpha_\ell(w) \exp[-c\, d(w, w_0)]\,,$$

where $d(\cdot, \cdot)$ is a string metric and $c > 0$.

New Work: Empirical Bayes Model

- Define $\alpha_\ell(w) =$ relative frequency of w in data for field ℓ.
- G_ℓ: empirical distribution for field ℓ.
- $W \sim F_\ell(w_0)$:

$$P(W = w) \propto \alpha_\ell(w) \exp[-c\, d(w, w_0)]\,,$$

where $d(\cdot, \cdot)$ is a string metric and $c > 0$.

$$X_{ij\ell} \mid \lambda_{ij},\ Y_{\lambda_{ij}\ell},\ z_{ij\ell} \sim \begin{cases} \delta(Y_{\lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 0 \\ F_\ell(Y_{\lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 1 \text{ and } \ell \le p_s \\ G_\ell & \text{if } z_{ij\ell} = 1 \text{ and } \ell > p_s \end{cases}$$

$$\begin{aligned} Y_{j'\ell} &\sim G_\ell \\ z_{ij\ell} \mid \beta_{i\ell} &\sim \text{Bernoulli}(\beta_{i\ell}) \\ \beta_{i\ell} &\sim \text{Beta}(a, b) \\ \lambda_{ij} &\sim \text{DiscreteUniform}(1, \ldots, N). \end{aligned}$$

Simulation Studies

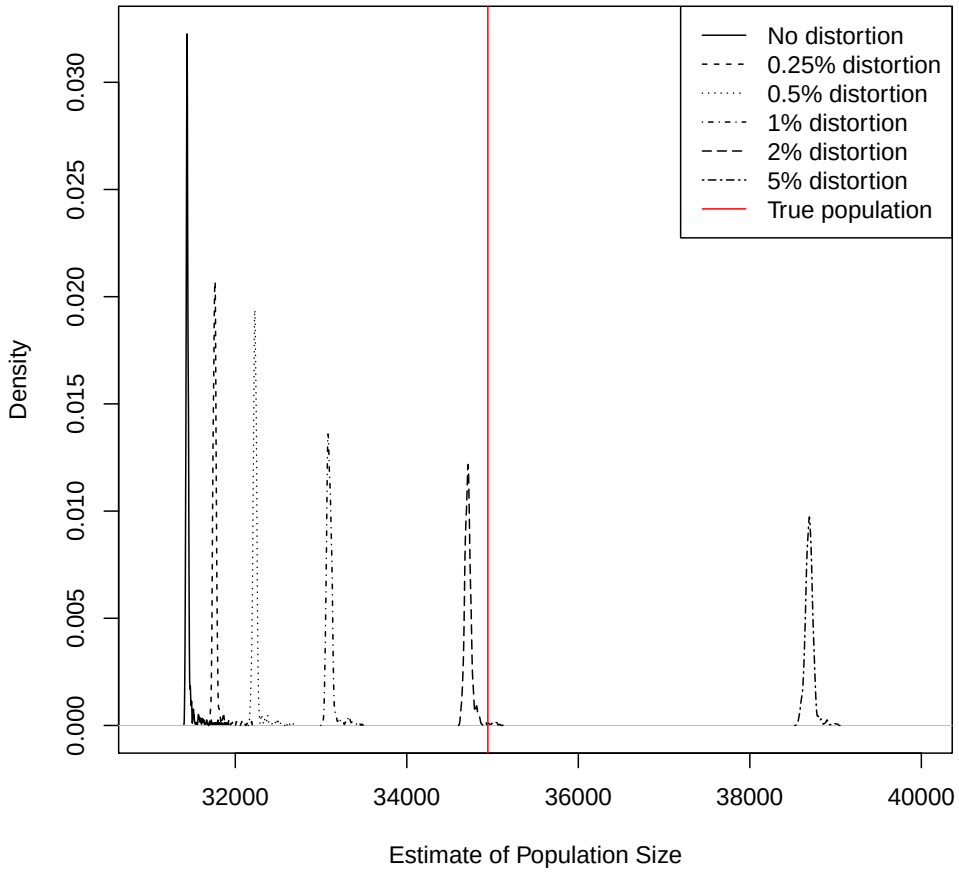


Figure : Posterior density estimates for 6 levels of distortion compared to ground truth (in red).

Simulation Studies

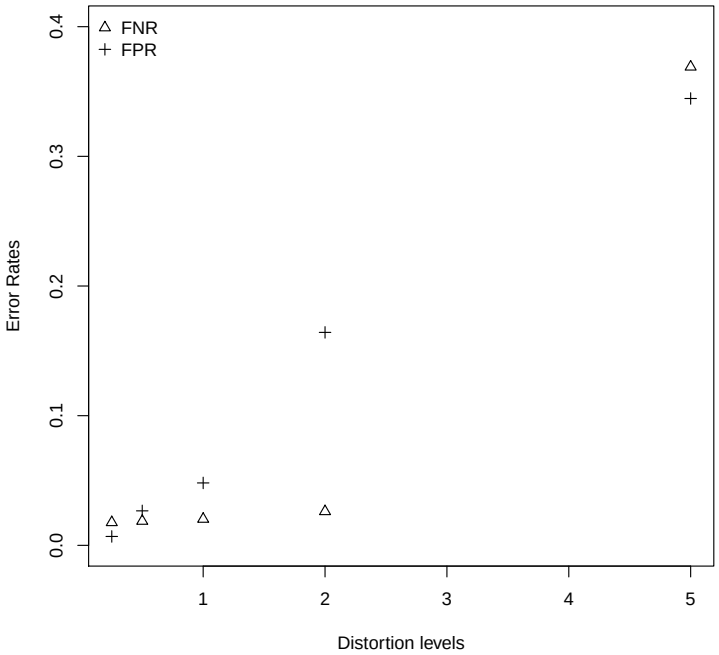


Figure : FNR and FPR plotted against 5 levels of distortion, where the former (plusses) shows near linear relationship and latter shows exponential one

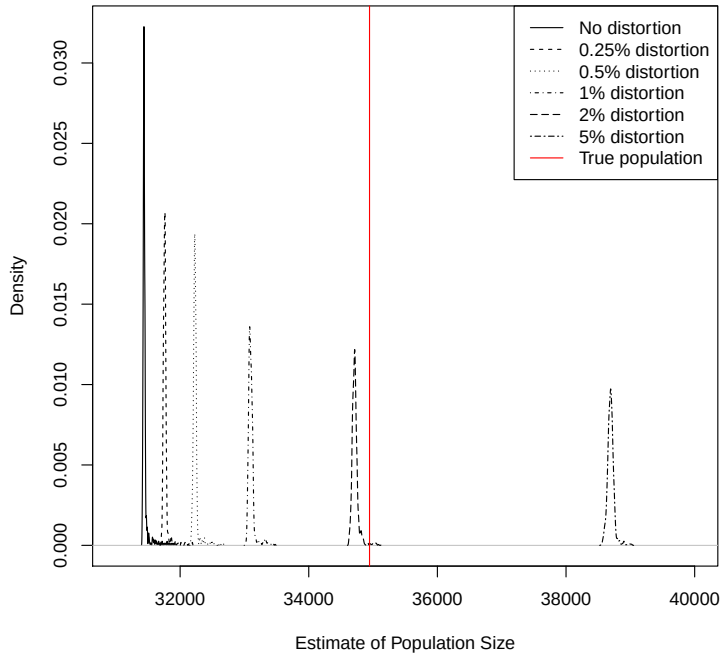
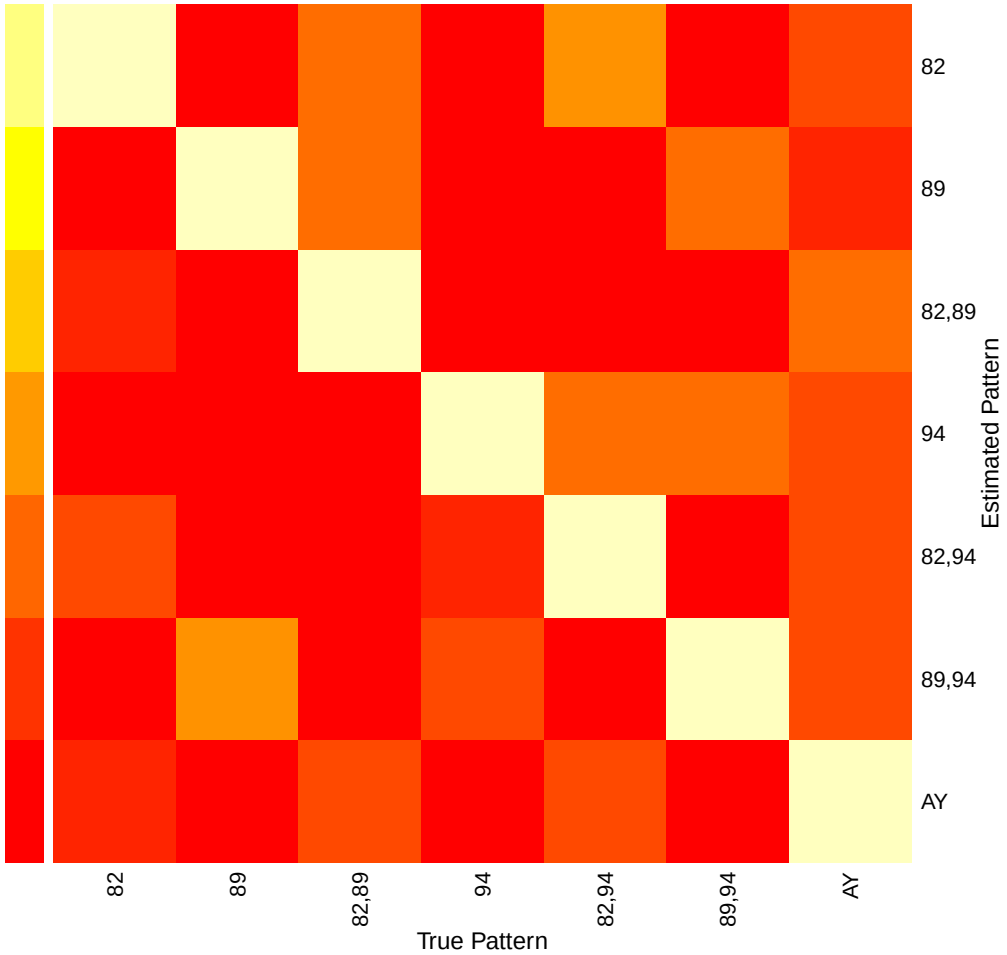


Figure : Posterior density estimates for 6 levels of distortion compared to ground truth (in red).

Confusion Matrix for NLTCS



Why Quantification of Uncertainty?

- Needed in subsequent analyses.
  - Capture recapture (Estimating those not in sample).
  - Small area estimation (Violence rates in small domains in Syria).
- Uncertainty quantification is automatic under the Bayesian paradigm.

Extensions to Strings

How does SMERED do comparing Stephen Fienberg to Steve Feinberg?

- The model of Steorts et al. (2013) was only built for categorical data and not string data.
- Steorts (2014) extends SMERED to an EB model, where the fields are strings and fields are dependent.

Why Split-Merge?

- Split-Merge is advantageous for high dimensional parameter spaces.
- Drawback of reversible jump: hard to design acceptable proposals and evaluate the Metropolis acceptance probabilities.

How many pairwise links end up being ignored due to requiring the "shared MPMMS" construction?

- Suppose that we only have one shared MPMMS. Then none are ignored.
- However, suppose that we have more than one shared MPMMS.
  - By construction this implies that each posterior probability < 0.5.
  - In general, we could have any finite number of shared MPMMSs, however, one will be most probable or will tie.
  - Hence, the number of pair-wise links that are ignored could be few or at most "very large" if they're all improbable.
  - This is application specific.

Algorithm 1: Split and MErge REcord linkage and Deduplication (SMERED)

Data: $\mathbf{X}$ and hyperparameters
Initialize the unknown parameters $\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{y}, \mathbf{z}$, and $\boldsymbol{\Lambda}$.

for $i \leftarrow 1$ **to** S_G **do**
  for $j \leftarrow 1$ **to** S_M **do**
    for $t \leftarrow 1$ **to** S_T **do**
      Draw records R_1 and R_2 uniformly and independently at random.
      if R_1 *and* R_2 *refer to the same individual* **then**
        propose splitting that individual, shifting $\boldsymbol{\Lambda}$ to $\boldsymbol{\Lambda}'$
      endif
      else
        propose merging individuals R_1 and R_2 refer to, shifting $\boldsymbol{\Lambda}$ to $\boldsymbol{\Lambda}'$
      endif
      $r \leftarrow \min\left\{1, \frac{\pi(\boldsymbol{\Lambda}', \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta} | \mathbf{x})}{\pi(\boldsymbol{\Lambda}, \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta} | \mathbf{x})}\right\}$
      Resample $\boldsymbol{\Lambda}$: accept proposed $\boldsymbol{\Lambda}'$ with probability r, otherwise reject
    end
    Resample $\mathbf{y}$ and $\mathbf{z}$.
  end
  Resample $\boldsymbol{\theta}, \boldsymbol{\beta}$.
end

return $\boldsymbol{\theta}|\mathbf{X}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{y}|\mathbf{X}, \mathbf{z}|\mathbf{X}$, *and* $\boldsymbol{\Lambda}|\mathbf{X}$.

Let $m = 1, \ldots, M_\ell$ index the possible categories of the ℓth field. Simplifying, we find that the joint posterior is

$$\begin{aligned}
&\pi(\mathbf{\Lambda}, \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta} \mid \mathbf{x}) \\
&\propto \prod_{i=1}^{k} \prod_{j=1}^{n_i} \prod_{\ell=1}^{p} \prod_{m=1}^{M_\ell} \left[(1 - z_{ij\ell}) \delta_{y_{\mathbf{\Lambda}_{ij}\ell m}}(x_{ij\ell}) \; + \; z_{ij\ell} \theta_{\ell m}^{I(x_{ij\ell}=m)} \right] \\
&\quad \times \prod_{\ell=1}^{p} \prod_{m=1}^{M_\ell} \theta_{\ell m}^{\mu_{\ell m} + \sum_{j'=1}^{N} I(y_{j'\ell}=m)} \\
&\quad \times \prod_{\ell=1}^{p} \beta_\ell^{a_\ell - 1 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} z_{ij\ell}} (1 - \beta_\ell)^{b_\ell - 1 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (1 - z_{ij\ell})}.
\end{aligned}$$

- In a large database setting, the EB method may still be slow.
- Consider novel blocking methods that seek to reduce the false negative errors.
- Borrow from the hashing literature.
- Goal: put similar records in the same bucket or block: locality sensitive hashing.
- Apply to a noisy database of $\approx$ 300,000 death records from Syria.

Ventura, Steorts, Nugent, and Furnish (2014).

1. Relies on domain knowledge to pick out the fields that are rarely error-free.
2. Can leave blocks so large that record linkage within blocks is still computationally infeasible.
3. Primary issue with blocking is that the scheme is not adapative.

- A *hash function* takes objects and maps them to integers, such that dissimilar objects are mapped far apart.
- LSH uses special hash functions to ensure that similar objects are put close to each other (with high probability).
- Apply LSH methods to a record.
  - Go from a high-dimensional object to a low-dimensional signature, where similar records have nearby signatures.
  - Low-dimensional signature vectors can be divided into blocks, with a high probability that all records mapped to the same bin are similar.
  - Hope that not too many similar records fall into different bins.

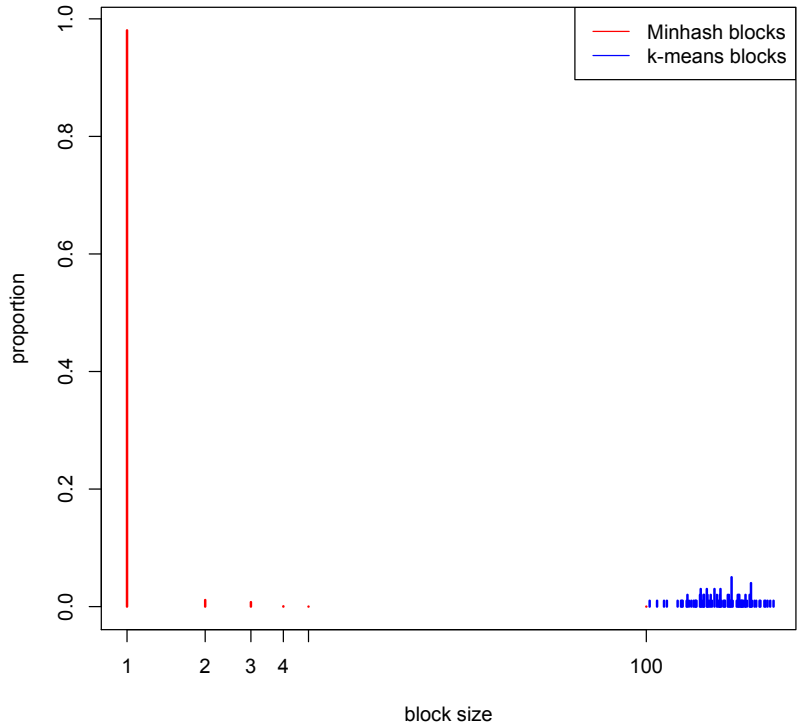


Figure : Bin distribution from using KLSH with an average bin size of 100 records per block and a random permutation of 20 (blue) contrasted with the bin distribution from using TLSH maximum bin size set to 200. KLSH results in more uniform distribution with equal sized blocks, whereas TLSH has one large block (size 100), and many undesirable singleton blocks (30,000 of these).

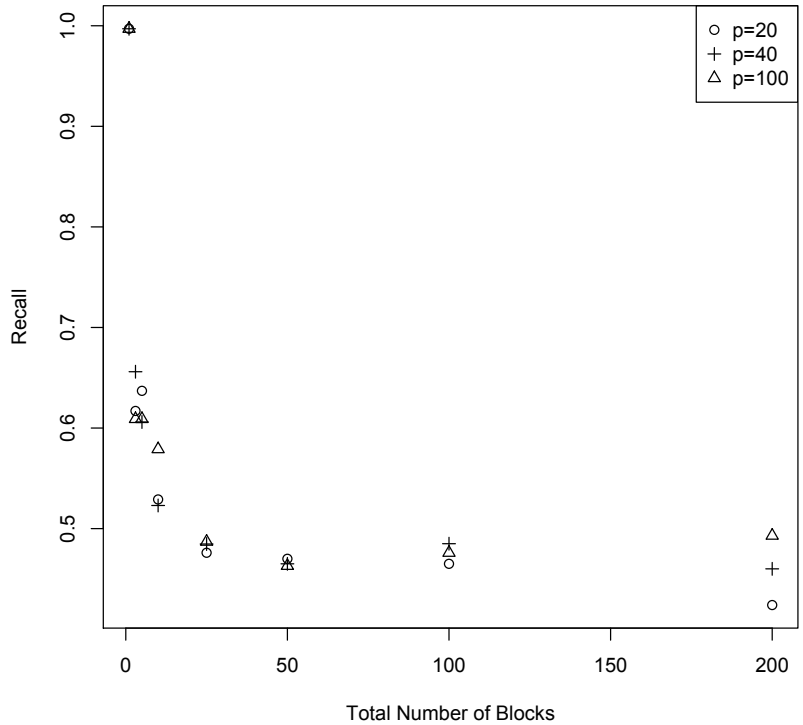


Figure : Illustrating the improvement to recall as the average number of records per bins decreases, with the number of random permutations p taken as 20, 40, and 100. We note that as p increases, the recall does not increase. The computational runtime of the blocking remains the same at 85s for each run with a p of 20.

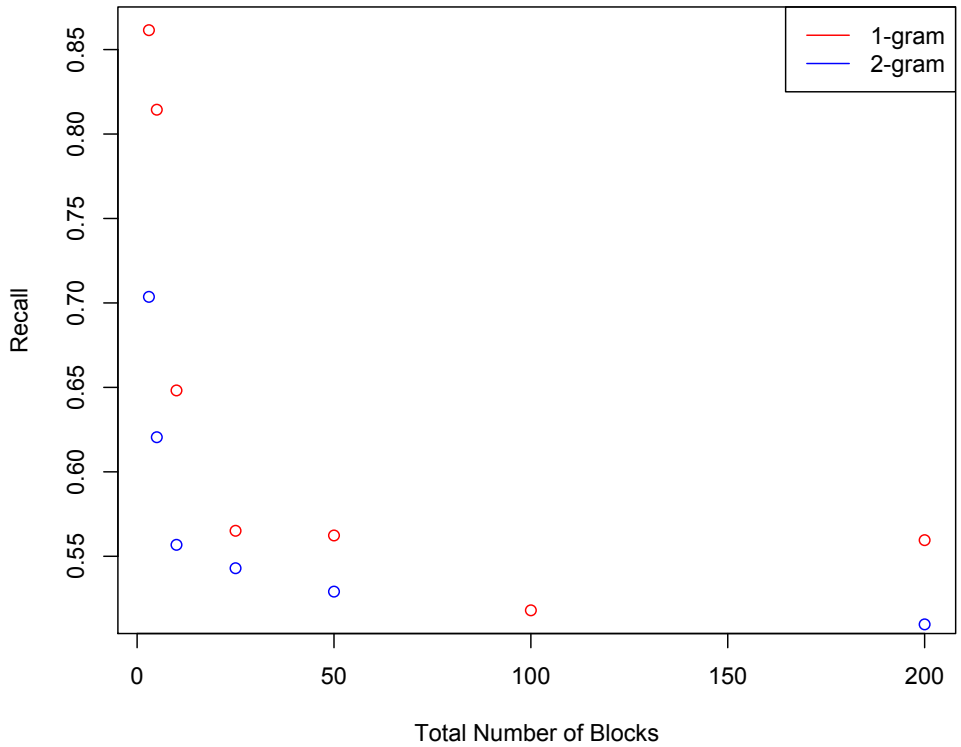


Figure : Illustrating the improvement to LSH when shingling each field attribute instead of treating the record as a bag of words. We use a 1- and 2-gram shingling LSH.

Let $X_1, X_2, \ldots$ be positive integer-valued random variables ("cluster assignment indices") such that $\sum_{i=1}^{\infty} \mathbf{1}_{\{m\}}(X_i) \leq k$ for all $m \geq 1$ (i.e., the mth cluster contains at most k variables).

Claim: $X_1, X_2, \ldots$ are not exchangeable.

Proof: Suppose $X_1, X_2, \ldots$ are exchangeable $\implies$ $X_1, X_2, \ldots$ have the same marginal distribution. $\exists$ positive integer m such that $P(X_1 = m) > 0$. Then

$$k = E(k) \geq E\left[\sum_{i=1}^{\infty} \mathbf{1}_{\{m\}}(X_i)\right] = \sum_{i=1}^{\infty} E\left[\mathbf{1}_{\{m\}}(X_i)\right] = \sum_{i=1}^{\infty} P(X_i = m) = \sum_{i=1}^{\infty} P(X_1 = m) = \infty,$$

a contradiction.[1]

[1] (Interchange of summation and integration above is justified by the nonnegativity of the summands/integrands by Fubini's Theorem (Hewitt and Stromberg).

Time Complexity

- SMERED is linear in both $N_{\max}$ and MCMC iterations.
- Let $M = \frac{1}{p}\sum_{\ell=1}^{p} M_\ell$ as the average number of possible values per field ($M \geq 1$).

- For S_G iterations of the Gibbs sampler, the algorithm is order $O(pMN_{\max}S_G S_M)$.
- If (p and M) $<<$ $N_{\max}$, the runtime is $O(N_{\max}S_G S_M)$.