# Entity Resolution with Societal Impacts in Statistical Machine Learning

Beidi Chen and Rebecca C. Steorts

Rice University, Department of Computer Science
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

H R D A G Human Rights Data Analysis Group
everybody counts.

February 8, 2017

# Summary

# Summary

- Entity resolution: merging large, noisy databases.

# Summary

- Entity resolution: merging large, noisy databases.
  - Medical data, official statistics, genetic data, human rights violations, and more. . . .

# Summary

- Entity resolution: merging large, noisy databases.
  - Medical data, official statistics, genetic data, human rights violations, and more. . . .
- In human rights, and more specifically in studies of conflict violence, we rarely have access to complete data.

# Summary

- Entity resolution: merging large, noisy databases.
    - Medical data, official statistics, genetic data, human rights violations, and more. . . .
- In human rights, and more specifically in studies of conflict violence, we rarely have access to complete data.
    - Instead, we have snapshots of violence and killings.

# Summary

- Entity resolution: merging large, noisy databases.
  - Medical data, official statistics, genetic data, human rights violations, and more. . . .
- In human rights, and more specifically in studies of conflict violence, we rarely have access to complete data.
  - Instead, we have snapshots of violence and killings.
- Seek important questions about human rights killings that use sound statistical and machine learning techniques.

# Summary

- Entity resolution: merging large, noisy databases.
  - Medical data, official statistics, genetic data, human rights violations, and more. . . .
- In human rights, and more specifically in studies of conflict violence, we rarely have access to complete data.
  - Instead, we have snapshots of violence and killings.
- Seek important questions about human rights killings that use sound statistical and machine learning techniques.
  - We wish to quantify the number of documented identifiable deaths in the Syrian conflict and quantify a standard error.

# Summary

- Entity resolution: merging large, noisy databases.
  - Medical data, official statistics, genetic data, human rights violations, and more. . . .
- In human rights, and more specifically in studies of conflict violence, we rarely have access to complete data.
  - Instead, we have snapshots of violence and killings.
- Seek important questions about human rights killings that use sound statistical and machine learning techniques.
  - We wish to quantify the number of documented identifiable deaths in the Syrian conflict and quantify a standard error.
  - We propose a statistical learning approach that is much less than quadratic, allowing computational efficiency.

# Summary

- Entity resolution: merging large, noisy databases.
    - Medical data, official statistics, genetic data, human rights violations, and more. . . .
- In human rights, and more specifically in studies of conflict violence, we rarely have access to complete data.
    - Instead, we have snapshots of violence and killings.
- Seek important questions about human rights killings that use sound statistical and machine learning techniques.
    - We wish to quantify the number of documented identifiable deaths in the Syrian conflict and quantify a standard error.
    - We propose a statistical learning approach that is much less than quadratic, allowing computational efficiency.
    - We do not make any distributional assumptions about the data generating process.

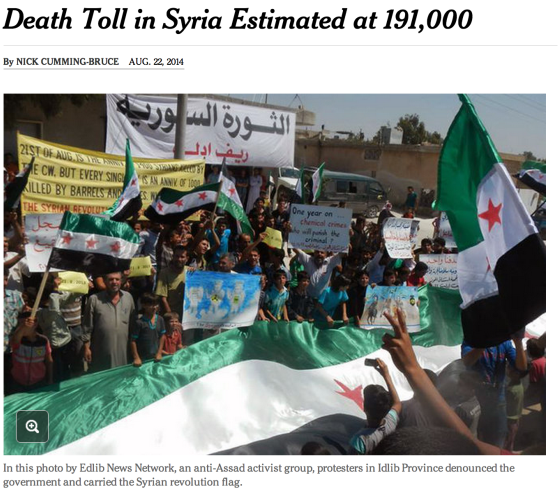
# Summary

- Entity resolution: merging large, noisy databases.
    - Medical data, official statistics, genetic data, human rights violations, and more. . . .
- In human rights, and more specifically in studies of conflict violence, we rarely have access to complete data.
    - Instead, we have snapshots of violence and killings.
- Seek important questions about human rights killings that use sound statistical and machine learning techniques.
    - We wish to quantify the number of documented identifiable deaths in the Syrian conflict and quantify a standard error.
    - We propose a statistical learning approach that is much less than quadratic, allowing computational efficiency.
    - We do not make any distributional assumptions about the data generating process.
    - Our proposed estimator is unbiased and has provably low variance compared to the current literature.

# Summary

- Entity resolution: merging large, noisy databases.
  - Medical data, official statistics, genetic data, human rights violations, and more. . . .
- In human rights, and more specifically in studies of conflict violence, we rarely have access to complete data.
  - Instead, we have snapshots of violence and killings.
- Seek important questions about human rights killings that use sound statistical and machine learning techniques.
  - We wish to quantify the number of documented identifiable deaths in the Syrian conflict and quantify a standard error.
  - We propose a statistical learning approach that is much less than quadratic, allowing computational efficiency.
  - We do not make any distributional assumptions about the data generating process.
  - Our proposed estimator is unbiased and has provably low variance compared to the current literature.
  - We illustrate our methodology on a subset of the Syrian conflict, which closely matches that from the 2014 Human Rights Data Analysis Group (HRDAG).

# Do we already know the answer?



*Death Toll in Syria Estimated at 191,000*

By NICK CUMMING-BRUCE AUG. 22, 2014

In this photo by Edlib News Network, an anti-Assad activist group, protesters in Idlib Province denounced the government and carried the Syrian revolution flag.

We build off of essential work that has been established by the Human Rights Data Analysis Group (HRDAG).

- We start with the raw data.
- We have four data sources, where the pattern over time is about the same (March 2011 — April 2014).

[Megan Price, Anita Gohdes, and Patrick Ball (2014)]

*How do we attempt estimating the number of documented identifiable deaths in Syria since March 2011?*

# Documented, Identifiable Victims

What resources we have:

- Human Rights Data Analysis Group collected 300,000 death records from Syria, which was published in a 2014 report with the United Nations.
- Records reported from four human rights sources.
- Field attributes: Full Arabic name, date of death (DOD), governorate, other less reliable ones.
- 40,000 record pairs labeled as match/non-match.

# HRDAG's August 2014 Release

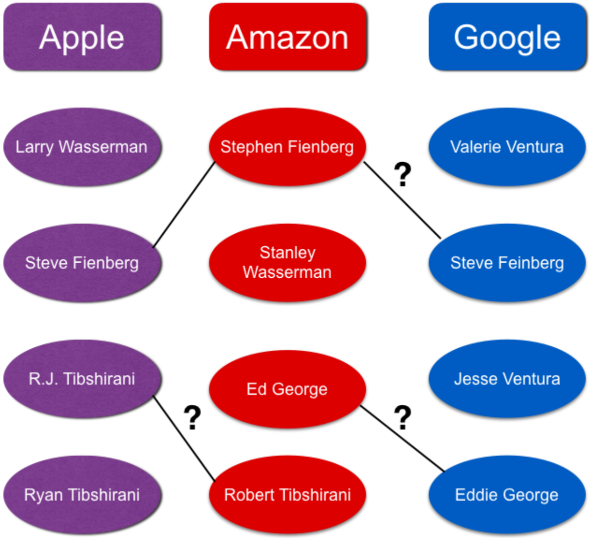
- Estimated 191,000 documented, identifiable deaths.
- Used hand matching (five people).

1. How reliable is the hand matching?
2. Standard error.
3. Scalability and cost.

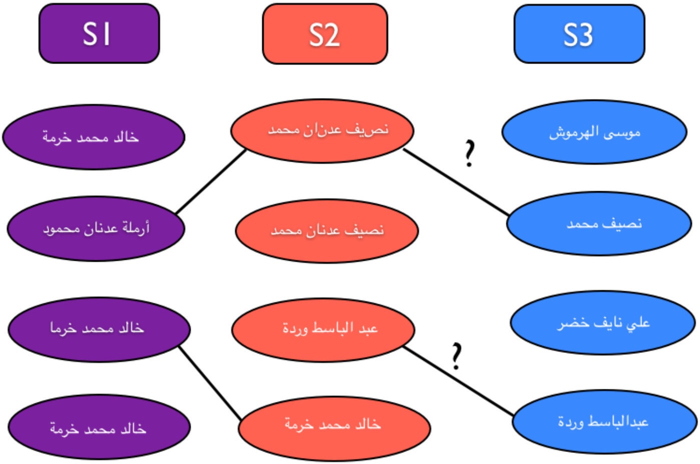
How can we improve on this approach?

*Entity resolution (record linkage or de-duplication) joins multiple data sets removes duplicate entities often in the absence of a unique identifier.*

# The record linkage graph



# The Syrian record linkage graph



S1

S2

S3

خالد محمد خرمة

نص‌يف عدن‌ان محمد

موسى الهرموش

?

أرملة عدنان محمود

نصيفة عدنان محمد

نصيف محمد

خالد محمد خرما

عبد الباسط وردة

علي نايف خضر

?

خالد محمد خرمة

خالد محمد خرمة

عبدالباسط وردة

# Entity Resolution

*Why is entity resolution difficult?*

# Goals of entity resolution

Suppose that we have a total of $M$ records in $D$ data sets.

1. We seek models that are much less than $O(M^2)$ (quadratic).
2. We seek models that are reliable, accurate, fit the data well, and account for the uncertainty of the model.

# Goals of entity resolution

Suppose that we have a total of $M$ records in $D$ data sets.

1. We seek models that are much less than $O(M^2)$ (quadratic).
2. We seek models that are reliable, accurate, fit the data well, and account for the uncertainty of the model.

These two goals fundamentally go against one another, making entity resolution a very challenging problem.

# Goals of entity resolution

Suppose that we have a total of $M$ records in $D$ data sets.

1. We seek models that are much less than $O(M^2)$ (quadratic).
2. We seek models that are reliable, accurate, fit the data well, and account for the uncertainty of the model.

These two goals fundamentally go against one another, making entity resolution a very challenging problem.

In order to solve the problem at hand, we will solve a slightly easier problem, where we simply provide an estimate and standard error of the number of documented identifiable deaths.

# Goals of entity resolution

Suppose that we have a total of $M$ records in $D$ data sets.

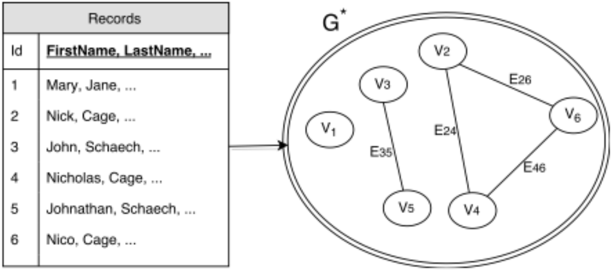
1. We seek models that are much less than $O(M^2)$ (quadratic).
2. We seek models that are reliable, accurate, fit the data well, and account for the uncertainty of the model.

These two goals fundamentally go against one another, making entity resolution a very challenging problem.

In order to solve the problem at hand, we will solve a slightly easier problem, where we simply provide an estimate and standard error of the number of documented identifiable deaths.

We refer to this subtask of entity resolution as unique entity estimation.

# The Linkage Graph



| Records | |
|---|---|
| Id | **FirstName, LastName, ...** |
| 1 | Mary, Jane, ... |
| 2 | Nick, Cage, ... |
| 3 | John, Schaech, ... |
| 4 | Nicholas, Cage, ... |
| 5 | Johnathan, Schaech, ... |
| 6 | Nico, Cage, ... |

Figure: Left: Records contained in *D* datasets. Right: The unlinked data set and relationship of records viewed as a graphical model.

We don't know the edges. How do we go about estimating the probability of an edge or non-edge?

# Desiderata

1. The estimation cost should be significantly less than quadratic ($O(M^2)$).
2. To ensure accountability regarding estimating the unique number of documented identifiable victims in the Syrian conflict, it is essential to understand the statistical properties of any proposed estimator.
3. In most real entity resolution tasks, duplicated data can occur with arbitrarily large changes including missing information, which we observe in the Syrian data set, and standard modeling assumptions may not hold due to the noise inherent in the data. Due to this, we prefer not to make strong modeling assumptions regarding the data generation process.

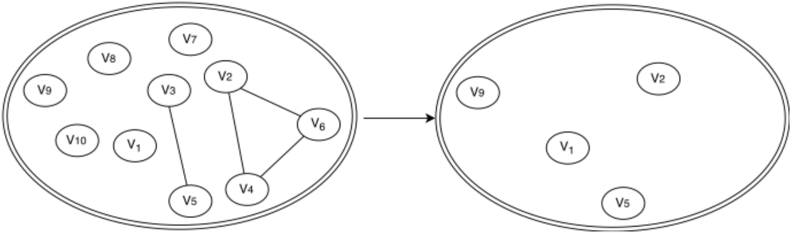
# Related Literature: Random Sampling

There are two methods based upon random sampling that do satisfy the above requirements, but they are sub-optimal.

1. Frank (1978) proposed sampling a large enough subgraph to estimate the total number of connected components based on the properties of the sub-sampled subgraph.
2. Chazelle, Rubinfeld, and Trevisan (2005) proposed finding connected components with high probability by sampling random vertices and then visiting their associated components using breadth-first search (BFS).

# Related Literature: Random Sampling

- One major issue with random sampling is that most sampled pairs are unlikely to be matches (no edge) providing nearly no information as the underlying graph is generally very sparse in practice.
- In addition, such methods often lead to a high variance of the resulting estimator.
- Finally, random sampling requires a threshold $t$ (fixed budget), which implies $t$ pairwise comparisons.



# Adaptive Sampling

- Here, we sample based on the similarity of the vertex pair.
- We assume that duplicates are likely to be similar in their textual attributes.
- We do not assume any fixed threshold, making the method robust and computationally effient.

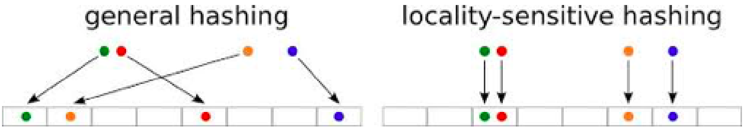
Does there exist a random sampling method based on similarity with computational complexity less than sub-quadratic?

Yes! (It's a variant of locality sensitive hashing (LSH).

# Locality Sensitive Hashing (LSH)

Hash functions are locality-sensitive if for any random hash function $h(\cdot)$ and for any pairs of data points (records) $x$ and $y$ we have the following:

1. Pr(h(x) = h(y)) is "high" if x is close to y.
2. Pr(h(x) = h(y)) is "low" if x if far from y.



As described earlier, we need to define a type of similarity and a type of dimension reduction.

We will use Jaccard similarity and the minwise hash (Shrivastava and Li, 2014).

# Our Contributions

1. We formalize unique entity estimation as approximating the number of connected components in a graph with sub-quadratic $\ll O(M^2)$ computational time.
2. We then propose a general method that provides an estimate in sample (with standard errors).
3. Our proposal leverages locality sensitive hashing (LSH) in a novel way for the estimation process, with the required computational complexity that is less than quadratic.
4. Our proposed estimator is unbiased and has provably low variance compared to comparable approaches in the literature.
5. We estimate that the number of documented identifiable deaths for the Syrian conflict is 191,874, with standard deviation of 1,772, reported casualties, which is very close to the 2014 HRDAG estimate of 191,369.

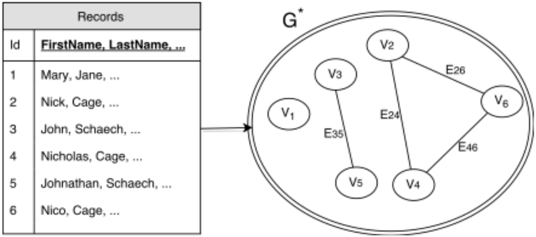
Chen, Shrivastava, **RCS** (2018), Minor Revision, AoAS
`https://arxiv.org/abs/1710.02690`,
Code Link

# Notation

- Let the total size of the data be $M$.
- Represent the data set by a graph $G^* = (E, V)$.
- Since we do not observe the edges of $G^*$ (the linkage), inferring whether there is an edge between two nodes can be costly, i.e., $O(M^2)$.
- Hence, one is constrained to probe a small set $\mathcal{S} \subset V \times V$ with $|\mathcal{S}| \ll O(M^2)$ of pairs and query if they have edges.
- The aim is to use the information about $\mathcal{S}$ to estimate the total number of connected components accurately.
- More precisely, given the partial graph $G' = \{V, E'\}$, where $E' = E \cap \mathcal{S}$, one wishes to estimate the connected components $n$ of $G^* = \{V, E\}$.

# Unique entity estimation

We approximate the number of connected components in the corresponding graph $G^*$, where the edges are not observed.



Records

| Id | FirstName, LastName, ... |
|---|---|
| 1 | Mary, Jane, ... |
| 2 | Nick, Cage, ... |
| 3 | John, Schaech, ... |
| 4 | Nicholas, Cage, ... |
| 5 | Johnathan, Schaech, ... |
| 6 | Nico, Cage, ... |

Figure: A toy example of mapping records to a graph, where vertices represent records and edges refer the relation between records.

# Fast Unique Entity Estimation

1. Generate pairs of records using locality sensitive hashing. This places similar pairs of records into the same bin in sub-quadratic time. Output is pairs of candidate records.
2. Query the candidate records in (1), and determine the match/non-match status either using (a) hand-matched data or (b) supervised learning algorithm. Now estimate $p$, the probability of an edge/non-edge. Next, we can form the graph $G^{'}$ from $p$ and count the number of connected components in $G^{'}$ using Breath First Search (BFS).
3. Now using the number of connected components, we can use our proposed method to estimate the number of unique entities in $G^{'}$.

# Fast Unique Entity Estimation

1. Generate pairs of records using locality sensitive hashing (LSH). This places similar pairs of records into the same bin in sub-quadratic time. Output is pairs of candidate records.

We use a linear variant of LSH (Shrivastava and Li, 2014) that has been applied to entity resolution to our knowledge for the first time successfully.

# Fast Unique Entity Estimation

1. Generate pairs of records using locality sensitive hashing (LSH). This places similar pairs of records into the same bin in sub-quadratic time. Output is pairs of candidate records.

We use a linear variant of LSH (Shrivastava and Li, 2014) that has been applied to entity resolution to our knowledge for the first time successfully.

Why does LSH work so well?

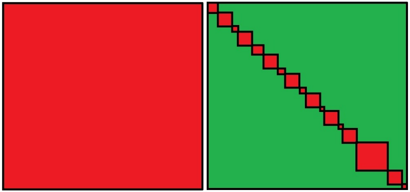


Figure: All-to-all record comparisons (left) versus partitioning records into blocks and comparing records only within each partition (right).

# Fast Unique Entity Estimation

2 Query the candidate records in (1), and determine the match/non-match status either using (a) hand-matched data or (b) supervised learning algorithm. Now estimate $p$, the probability of an edge/non-edge. Next, we can form the graph $G^{'}$ from $p$ and count the number of connected components in $G^{'}$ using BFS.

Using LSH, we reduce the record space greatly and have a new sample space $S$ of candidate records. Let $T_{match}$ represent the record pairs in our training set that are labeled as a match. Then an unbiased estimated of $p$ (the probability that any given correct pair is sampled) is

$$\frac{|T_{match} \cap S|}{T_{match}}. \tag{1}$$

Given this estimate of $p$, we can form $G^{'}$ and count the number of connected components using BFS.

# Fast Unique Entity Estimation

3. Now using the number of connected components, we can use our proposed method to estimate the number of unique entities in $G^{'}$.

Assuming we have an estimate of $p$, we are simply just solving the following system of equations that will map us from the unknown graph to the known graph (via $p$).

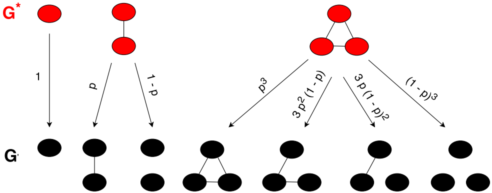


Figure: A general example illustrating the transformation and probabilities of connected components from $G^*$ to $G'$.

# Fast Unique Entity Estimation

Let $n'_i$ be the number of connected components of size $i$ in the observed graph G′.

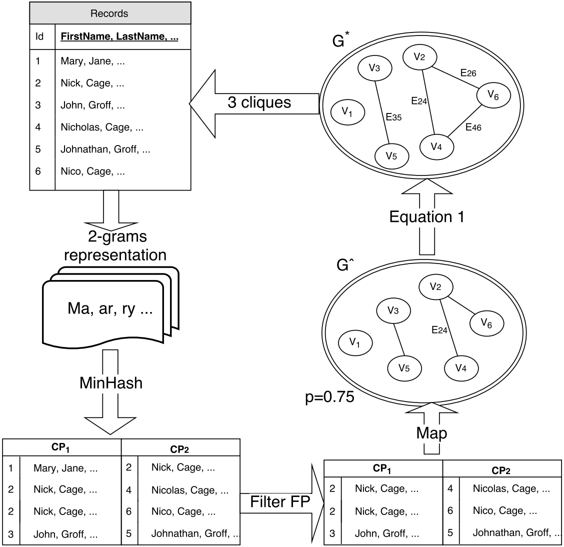
We propose an estimator and assume there exists an algorithm that adaptively samples a set of record pairs, in sub-quadratic time.

Our estimator, which we call the Locality Sensitive Hashing Estimator (LSHE) for the number of connected components is given by

$$\mathrm{LSHE} = n'_1 + n'_2 \cdot \frac{2p-1}{p} + n'_3 \cdot \frac{1 - 6 \cdot (1-p)^2 \cdot p}{p^2 \cdot (3-2p)} + \sum_{i=4}^{M} n'_i. \quad (2)$$

It is simple to show that this estimator is unbiased and has provably low variance.

# Fast Unique Entity Estimation



| Records | |
|---|---|
| Id | **FirstName, LastName, ...** |
| 1 | Mary, Jane, ... |
| 2 | Nick, Cage, ... |
| 3 | John, Groff, ... |
| 4 | Nicholas, Cage, ... |
| 5 | Johnathan, Groff, ... |
| 6 | Nico, Cage, ... |

| **CP$_1$** | | **CP2** | |
|---|---|---|---|
| 1 | Mary, Jane, ... | 2 | Nick, Cage, ... |
| 2 | Nick, Cage, ... | 4 | Nicolas, Cage, ... |
| 2 | Nick, Cage, ... | 6 | Nico, Cage, ... |
| 3 | John, Groff, ... | 5 | Johnathan, Groff, .. |

| **CP$_1$** | | **CP2** | |
|---|---|---|---|
| 2 | Nick, Cage, ... | 4 | Nicolas, Cage, ... |
| 2 | Nick, Cage, ... | 6 | Nico, Cage, ... |
| 3 | John, Groff, ... | 5 | Johnathan, Groff, .. |

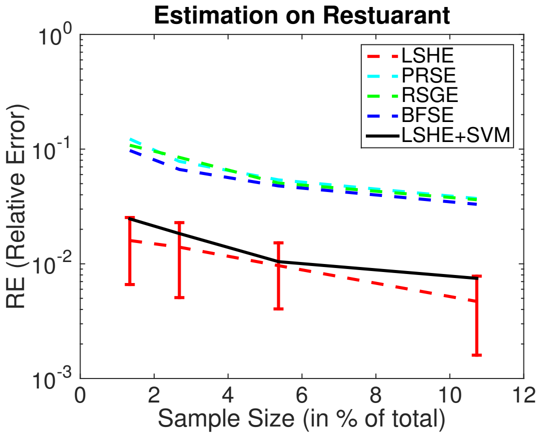
# Applications

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

Table: presents five important features of the four data sets. **Domain** reflects the variety of the data type we used in the experiments. **Size** is the number of total records respectively. **# Matching Pairs** shows how many pair of records point to the same entity in each data set. **# Attributes** represents the dimensionality of individual record. **# Entities** is the number of unique records.

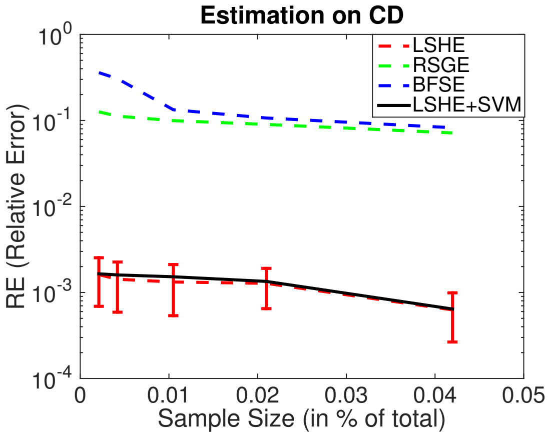
| **DBname** | **Domain** | **Size** | **# Matching Pairs** | **# Attributes** | **# Entities** |
|---|---|---|---|---|---|
| Restaurants | Restaurant Guide | 864 | 112 | 4 | 752 |
| CD | Music CDs | 9,763 | 299 | 106 | 9,508 |
| Voter | Registration Info | 324,074 | 70,359 | 6 | 255,447 |
| Syria | Death Records | 296,245 | N/A | 7 | N/A |

# Results



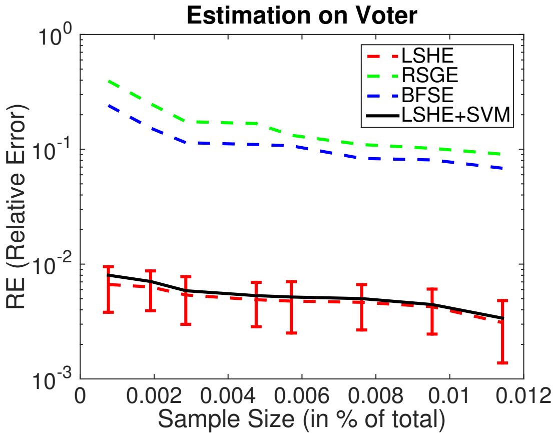
- Our methods: Red (LSHE) and black (LSHE + SVM)
- Comparisons: blue and green (random sampling)

# Results



- Our methods: Red (LSHE) and black (LSHE + SVM)
- Comparisons: blue and green (random sampling)

# Results



- Our methods: Red (LSHE) and black (LSHE + SVM)
- Comparisons: blue and green (random sampling)

# Application to Syrian Conflict

One main challenge with Syria is that the hand-labeled data is biased. A major goal is to also use this method for an automated labeling process.

- Since we have a reasonable size of manually labeled pairs, we train a support vector machine based upon these pairs, where we split data into a training and test set.
- We verify the labeling accuracy of the SVM's on the test set.
- We believe this to be reasonable given the accuracy on approaches for the three real data sets considered earlier.
- That is, when ground truth is noisy or biased, this method can be used as a proxy for ground truth measurements.

# Evaluation Methods

1. Pairs of data can be linked in both the handmatched training data (which we refer to as “truth”) and under the estimated linked data. We refer to this situation as true positives (TP).
2. Pairs of data can be linked under the truth but not linked under the estimate, which are called false negatives (FN).
3. Pairs of data can be not linked under the truth but linked under the estimate, which are called false positives (FP).
4. Pairs of data can be not linked under the truth and also not linked under the estimate, which we refer to as true negatives (TN).

# Recall, Precision, Reduction Ratio

$$\text{Recall} = \frac{TP}{TP + FN} = 1 - FNR.$$

# Recall, Precision, Reduction Ratio

$$\text{Recall} = \frac{TP}{TP + FN} = 1 - FNR.$$

$$\text{Precision} = \frac{TP}{TP + FP} = 1 - FPR.$$

# Recall, Precision, Reduction Ratio

$$\text{Recall} = \frac{TP}{TP + FN} = 1 - FNR.$$

$$\text{Precision} = \frac{TP}{TP + FP} = 1 - FPR.$$

Reduction ratio (RR) measures the relative reduction of the comparison space from the de-duplication or hashing technique.

See Christen (2012), Steorts, Ventura, Sadinle, Fienberg (2014) for a formal definition.

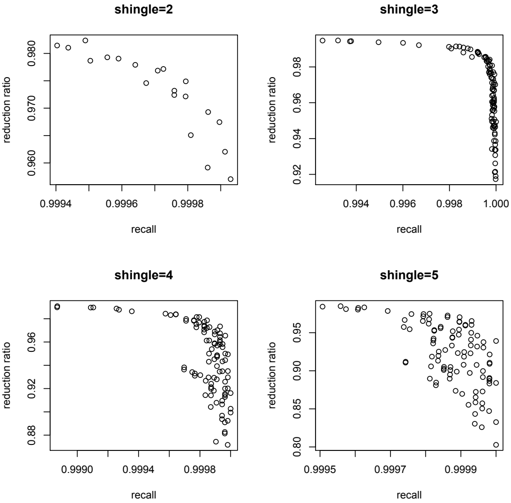
# Final Estimate

- Using our proposed methodology, used 917,577 sampled pairs and then used an SVM for classification of matches and non-matches.
- After looking at a sensitivity analysis of the tuning parameters for our proposed method, we report that there are 191,874 documented identifiable deaths, with standard deviation of 1,772, which is very close to HRDAG's estimate of 191,369 in Price et al. (2014).
- Furthermore, we also report the recall (false positives) = 0.83 and the precision (false negatives) = 0.99.
- Finally, one iteration of the method takes only 127 seconds! (Real time estimation is possible).

Questions? beka@stat.duke.edu

Thank you to Patrick Ball and Megan Price at the Human Rights Data Analysis Group for providing the data and for fruitful conversations that have constantly improved the work. This work would not have been possible without the encouragement and mentorship of Steve Fienberg.

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.


shingle=2
reduction ratio
0.960
0.970
0.980
0.9994
0.9996
0.9998
recall
shingle=3
reduction ratio
0.92
0.94
0.96
0.98
0.994
0.996
0.998
1.000
recall
shingle=4
reduction ratio
0.88
0.92
0.96
0.9990
0.9994
0.9998
recall
shingle=5
reduction ratio
0.80
0.85
0.90
0.95
0.9995
0.9997
0.9999
recall

# What is a hash function?

Find a hash function $h()$ such that

- if $\text{sim}(A, B)$ is high, then with high prob. $h(A) = h(B)$.
- if $\text{sim}(A, B)$ is low, then with high prob. $h(A) \neq h(B)$.

# Locality-Sensitive Hashing

- LSH tries to preserve similarity after dimension reduction.
  - What kind of similarity? ↔ What kind of dimension reduction?

# Empirical Bayes Model

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

# Empirical Bayes Model

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

# Empirical Bayes Model

- Define $\alpha_\ell(w)$ = relative frequency of $w$ in data for field $\ell$.
- $G_\ell$: empirical distribution for field $\ell$.
- $W \sim F_\ell(w_0)$: $P(W = w) \propto \alpha_\ell(w) \exp[-c\, d(w, w_0)]$, where $d(\cdot, \cdot)$ is a string metric and $c > 0$.

# Empirical Bayes Model

- Define $\alpha_\ell(w)$ = relative frequency of $w$ in data for field $\ell$.
- $G_\ell$: empirical distribution for field $\ell$.
- $W \sim F_\ell(w_0)$: $P(W = w) \propto \alpha_\ell(w) \exp[-c\, d(w, w_0)]$, where $d(\cdot,\cdot)$ is a string metric and $c > 0$.

$$\begin{aligned}
X_{ij\ell} \mid \lambda_{ij},\ Y_{\lambda_{ij}\ell},\ z_{ij\ell} &\sim \begin{cases} \delta(Y_{\lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 0 \\ F_\ell(Y_{\lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 1 \text{ and field } \ell \text{ is string-valued} \\ G_\ell & \text{if } z_{ij\ell} = 1 \text{ and field } \ell \text{ is categorical} \end{cases} \\
Y_{j'\ell} &\sim G_\ell \\
z_{ij\ell} \mid \beta_{i\ell} &\sim \text{Bernoulli}(\beta_{i\ell}) \\
\beta_{i\ell} &\sim \text{Beta}(a, b) \\
\lambda_{ij} &\sim \text{DiscreteUniform}(1, \ldots, N_{\max}), \qquad \text{where } N_{\max} = \sum_{i=1}^{k} n_i.
\end{aligned}$$

# Why Bayesian?

1. Latent entity for clustering allows us to handle many databases.
2. Method is general to many applications.
3. Scalability for categorical data is quite good.
4. We can provide a summary of the linkages using what we call the shared most probable maximal matching set.
5. The linkage structure is flexible – provides more than just links/non-links.

# Error Propagation for Estimating Death Counts

$$Var(N \mid \mathbf{X}) = Var_{\mathbf{\Lambda}|\mathbf{X}} E[N \mid \mathbf{\Lambda}]$$

# Error Propagation for Estimating Death Counts

$$Var(N \mid \mathbf{X}) = Var_{\mathbf{\Lambda}|\mathbf{X}} E[N \mid \mathbf{\Lambda}]$$

This estimate is only as good as the model that we are fitting!

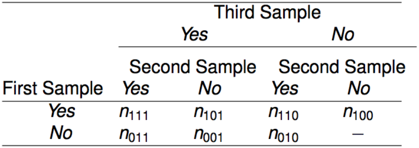
# Total Variance Decomposition

$$\begin{aligned} Var(N \mid \mathbf{X}) &= Var_{\mathbf{\Lambda}|\mathbf{X}} E[N \mid \mathbf{\Lambda}] && (3) \\ &= E_{\mathbf{\Lambda}|\mathbf{X}} \left\{ Var_{m|\mathbf{\Lambda}} E[N \mid \mathbf{\Lambda}, m] \right\} && (4) \\ &+ E_{\mathbf{\Lambda}|\mathbf{X}} \left\{ E_{m|\mathbf{\Lambda}} Var[N \mid \mathbf{\Lambda}, m] \right\} && (5) \end{aligned}$$

# Population sized estimation

- Target: size $N$ of a population (not sample).

$\boldsymbol{n} =$



| | Third Sample | | | |
|---|---|---|---|---|
| | *Yes* | | *No* | |
| | Second Sample | | Second Sample | |
| First Sample | *Yes* | *No* | *Yes* | *No* |
| *Yes* | $n_{111}$ | $n_{101}$ | $n_{110}$ | $n_{100}$ |
| *No* | $n_{011}$ | $n_{001}$ | $n_{010}$ | – |

- From these capture histories we can find $p_A(N \mid \boldsymbol{n})$.

# Population sized estimation

Uncertainty in $\boldsymbol{\Lambda}$ is captured by $p_L(\boldsymbol{\Lambda} \mid \mathbf{X})$.

$$\begin{aligned} p_C(N \mid \mathbf{X}) &\propto \sum_{\boldsymbol{\Lambda}} \text{linkage model likelihood} \times p(N, \boldsymbol{\Lambda}) \\ &= \sum_{\boldsymbol{\Lambda}} \text{linkage model likelihood} \times p_A(N \mid \boldsymbol{n}) p(\boldsymbol{\Lambda}). \end{aligned}$$