Entity Resolution:
Measuring and Reporting Quality

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

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

January 25, 2018

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

Motivations

In Syria, we have duplicated information regarding individuals who have died in the conflict.
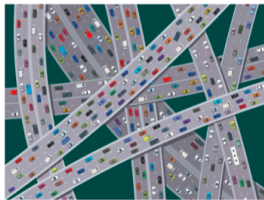
In the census, we have duplicated information of individuals that fill out census forms every 10 years.

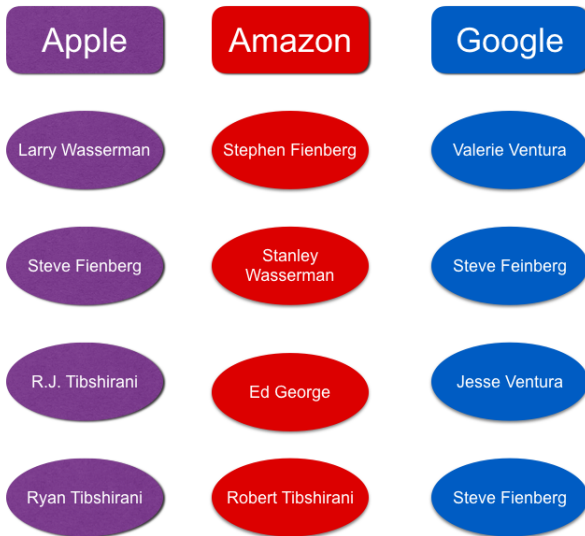Goal: Estimation of the sample size and associated standard errors.

AARIAN MARSHALL   TRANSPORTATION 05.11.17 8:30 AM

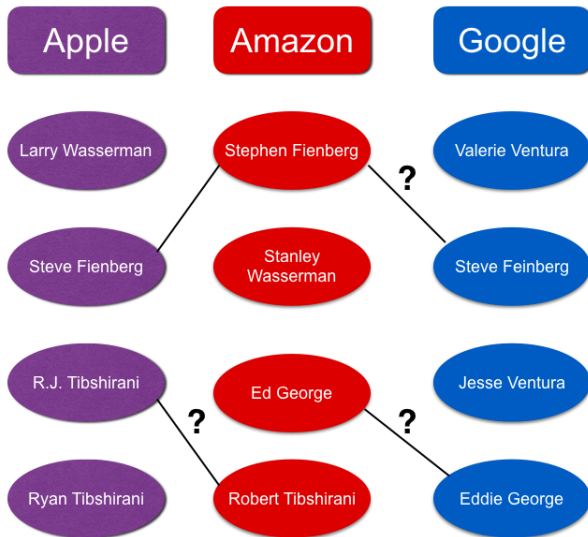## BAD NEWS FOR EVERYONE! THE 2020 CENSUS IS ALREADY IN TROUBLE

GETTY IMAGES

A graph with no edges

Apple	Amazon	Google
Larry Wasserman	Stephen Fienberg	Valerie Ventura
Steve Fienberg	Stanley Wasserman	Steve Feinberg
R.J. Tibshirani	Ed George	Jesse Ventura
Ryan Tibshirani	Robert Tibshirani	Steve Fienberg

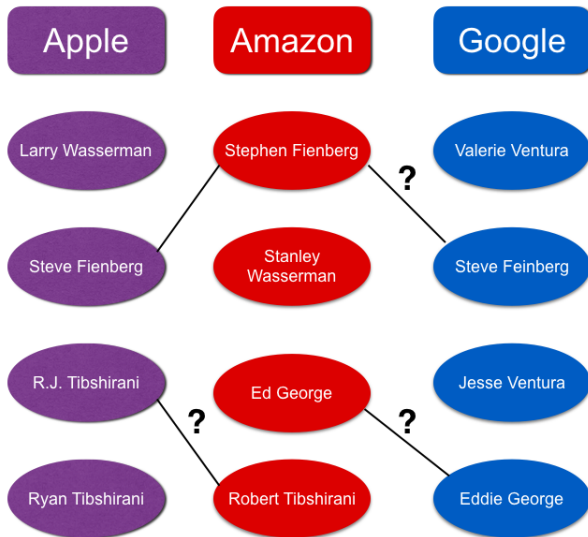# The record linkage graph

The node of Larry Wasserman

The node of Larry Wasserman

Larry Wasserman

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

The record linkage graph

Who's the real Steve Fienberg?

Steve Feinberg

240 Collins Drive
Pittsburgh, PA
50-54
412-793-3313

Who's the real Steve Fienberg?

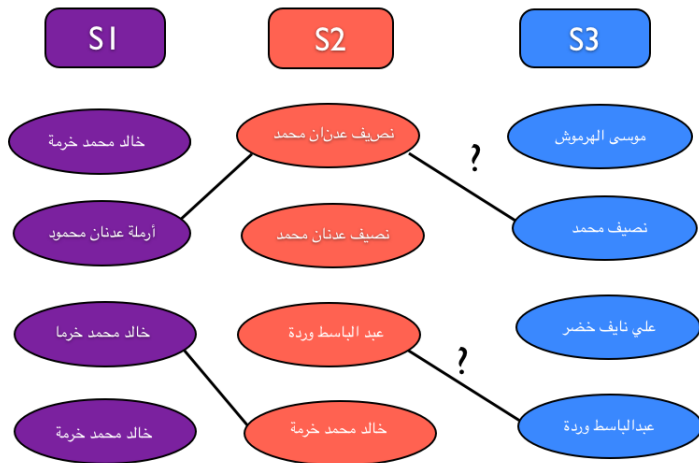240 Collins Drive
Pittsburgh, PA
50-54
412-793-3313

537 N Neville Street
Pittsburgh, PA 15213
65+
412-683-5599

These are clearly not the *same* Steve Fienberg!

Syrian Civil War

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 record linkage 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 record linkage a very challenging problem.

Depending on the motivating goal of a record linkage task, we approach it using either 1 or 2.

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.

In the rest of the talk, we will
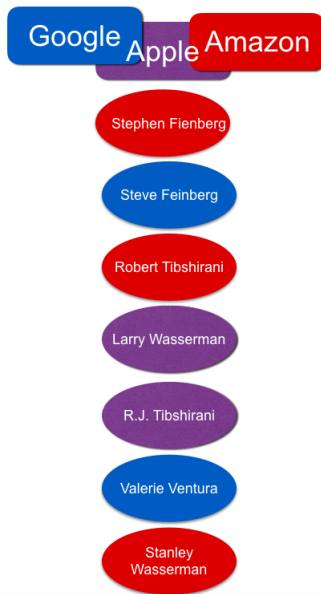
1. Review the literature.
2. Present Bayesian methods that satisfy 2.
3. And present a framework that satisfies both 1 and 2 (with a restriction).

Terminology

1. De-duplication
2. Record linkage
3. Blocking

De-duplication

De-duplication

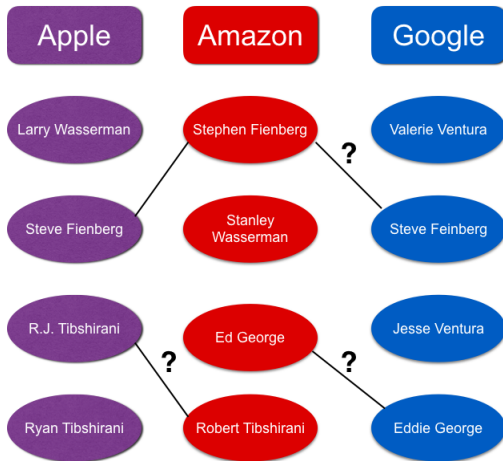Much of the literature can be grouped into the case of de-duplication.

Common examples from both academia and industry are the following:

logistic regression, random forests, support vector machines, Bayesian adaptive regression trees, and locality sensitive hashing.

The record linkage graph

Here, we call this record linkage, since we look at the record linkage uncertainty of the entire graphical structure.

Blocking

Often one performs blocking due to the fact that record linkage problems require a quadratic number of comparisons.
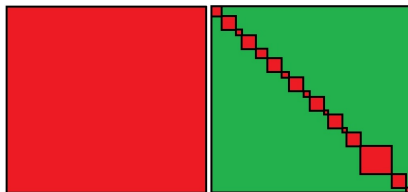
Figure: All-to-all record comparisons (left) versus partitioning records into blocks and comparing records only within each partition (right).

We will assume some method of blocking is embedded within a record linkage procedure.

Blocking

The most common method used for blocking is typically

1. deterministic blocking method
2. probabilistic blocking method

Examples include blocking on features (deterministic) or probabilistic types such as locality sensitive hashing.

See Christen (2012); Steorts, Ventura, Sadinle, Fienberg (2014); Chen, Shrivastava, Steorts (2017).

Common Methods for Entity Resolution

- Match on a unique identifier if it exists.
- Perform exact matching.
- Perform a likelihood ratio or hypothesis test.

[Newcombe (1959), Fellegi and Sunter (1969)].

Unique Identifier

Suppose that each feature has a unique identifier that we are sure is accurate, like social security number.

Then we can unique match records based on the unique identifier.

Problems occur this unique identifier is missing or has noise in it, etc.

Exact Matching

In exact matching, we compare all features. We decide if the record is a match if they agree on all features. Otherwise, we decide the record is a non-match.

Steve Feinberg	Stephen Fienberg
240 Collins Drive	537 N Neville Street
Pittsburgh, PA	Pittsburgh, PA 15213
50-54	65+
412-793-3313	412-683-5599

Why would this method be bad for evaluation purposes?

Fellegi and Sunter Method

- Newcombe (1959), Fellegi and Sunter (1969): two databases, all-to-all comparison of records.
- Neyman Pearson hypothesis test with a threshold t.
- If record i and j are above t, then we have a match.
- Otherwise, a non-match.

Fellegi and Sunter Method

- Computationally intractable.
- Transitivity not preserved.
- If 1 matches 2, and 2 matches 3, then 1 does NOT necessarily match 3.

Major limitations and major flaws of this method!
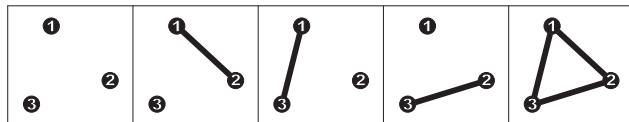
Evaluation Metrics

How do we evaluate performance of a particular record linkage method?

Evaluation Metrics

1. Recall
2. Precision
3. Reduction Ratio
4. Estimated Sample Size
5. Standard Error of Estimated Sample Size
6. Run Time
7. Robustness to Tuning Parameters
8. Do Supervised Methods Overfit the Data

Evaluation Metrics

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

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

Recall, Precision, Reduction Ratio

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

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

Recall, Precision, Reduction Ratio

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

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

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.

Other Evaluation Metrics

1. Estimated Sample Size
2. Standard Error of Estimated Sample Size
3. Run Time
4. Robustness to Tuning Parameters
5. Do Supervised Methods Overfit the Data

1. The estimated sample size and standard error must be defined for each method, but this is not difficult to do in practice.
2. Any method can be evaluated also for the run time, so one can gauge computationally costs.
3. Robustness of tuning parameters should be explored from a Bayesian and a frequentist perspective.
4. It's also essential to make sure that supervised methods do not overfit the data (see Steorts (2015)).

RLdata500 dataset

Let's look at an example on data that is available from the Record
Linkage package in R, where we compare many different methods
according to the evaluation metrics that we have laid out.

We will first describe the data set and then compare the following
methods in R:

1. blink
2. logistic regression
3. random forests
4. Bayesian adaptive regression trees

RLdata500 dataset

```
  fname_c1 fname_c2 lname_c1 lname_c2   by bm bd
1  CARSTEN    <NA>    MEIER     <NA> 1949  7 22
2     GERD    <NA>    BAUER     <NA> 1968  7 27
3   ROBERT    <NA> HARTMANN     <NA> 1930  4 30
4   STEFAN    <NA>    WOLFF     <NA> 1957  9  2
5     RALF    <NA>  KRUEGER     <NA> 1966  1 13
6  JUERGEN    <NA>   FRANKE     <NA> 1929  7  4
```
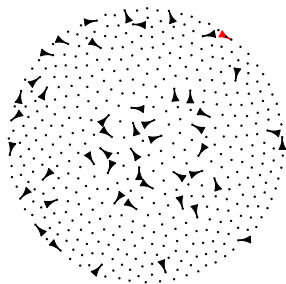
The RLdata500 data set consists of 500 records with 10 percent
duplication.
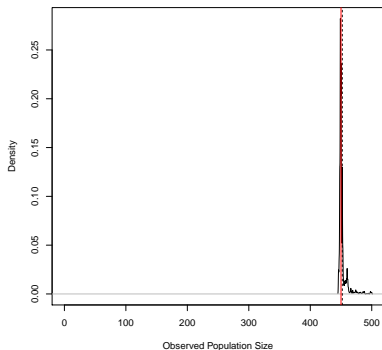
Figure: Posterior density for N in simulation study. The FNR and FPR: 0.04 and 0.02.

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

Procedure	FNR	FDR
blink (Steorts (2015))	0.02	0.08
Exact Matching	1	0
Near-Twin	1	0
BART (10% training)	0.10	0.16
BART (20% training)	0.07	0.11
BART (50% training)	0.03	0.04
BART (full data)	0.02	0
Random Forests (10% training)	0.05	0.15
Random Forests (20% training)	0.04	0.09
Random Forests (50% training)	0.02	0.06
Random Forests (full data)	0.04	0.06
Logistic Regression (10% training)	0.09	0.16
Logistic Regression (20% training)	0.06	0.07
Logistic Regression (50% training)	0.02	0.01
Logistic Regression (full data)	0.02	0

Table: False negative rate (FNR) and false discovery rate (FDR) for the proposed EB methodology and five other record linkage methods. For the supervised methods, we run 100 iterations of each one and average these such that overfitting is not occurring.

Robustness

How do we make sure a method is robust?

For a semi-supervised method, we want to make sure that it's robust to different choices of the training/test data and any tuning paramater(s).

For probabilistic and Bayesian methods, we want to make sure these methods are robust to choices of hyper-parameters and/or tuning parameters.

Robustness, computational time complexity, and sensitivity analysis can be further explored in Steorts (2015) and Steorts, Hall, Fienberg (2016).

What Record Linkage Needs

1. Workshops and Forums
2. Open Source Software
3. Reproducibility and Fairness
4. Evaluation Metrics and Comparisons
5. Transparency
6. Ethical Use of Data
7. Proper Privacy Protections of Sensitive Data

Some Key Papers

1. Christen (2012); Steorts, Ventura, Sadinle, Fienberg (2016) [Fair Comparisons]

2. Sadinle (2014), Steorts, Hall, Fienberg (2016) [Fair Comparisons]

3. Vatsalan, Christen, O'Keefe, Verykios (2014) [Evaluations of Record Linkage and Privacy]

4. Victoria Stodden and David Dunson [Reproducibility and Fairness]

5. Chen, Shrivastava, Steorts (2018) [Reproducibility, Fairness, and Evaluations]

6. David Banks, Megan Price, Patrick Ball [Ethical Use of Data with respect to Human Rights]

Thank you! Questions?

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.

Contact: beka@stat.duke.edu
Webpage: resteorts.github.io
(All papers and software can be found above).

Let's turn back to our motivating application of the US Census or the Syrian conflict. How would we estimate the sample and standard error in practice?

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.

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.

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 documented identifiable deaths.

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 documented identifiable deaths.

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

Our Contributions

1. We formalize unique entity estimation as approximating the number of connected components in a graph with sub-quadratic computational time.

2. Our proposed methodology makes no assumptions regarding the generating process of the records and gives an estimate in sample (with standard errors).

3. Our proposal leverages locality sensitive hashing (LSH) in a novel way for the estimation process, where our estimator is unbiased and has provably low variance compared to random sampling based approaches.

4. We apply our method to official statistics data, music data, food data, and a subset of the ongoing conflict in Syria.

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

Fast Unique Entity Estimation
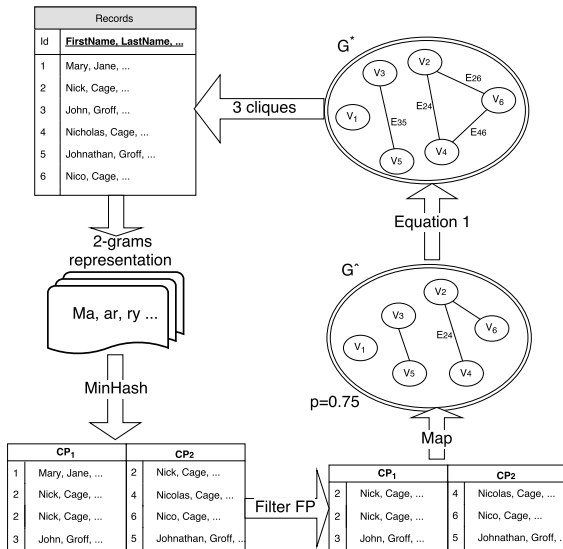
Applications

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
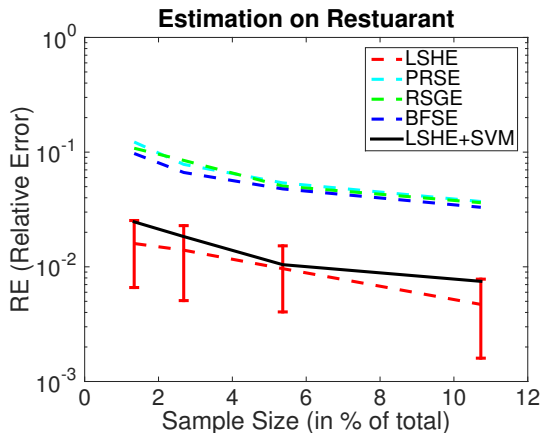
Estimation on Restuarant

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

Results

Estimation on CD

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

Results

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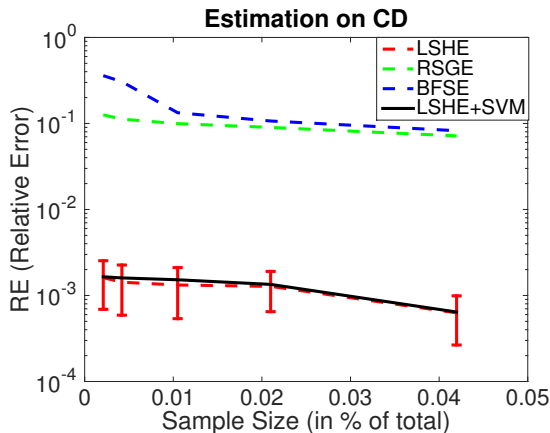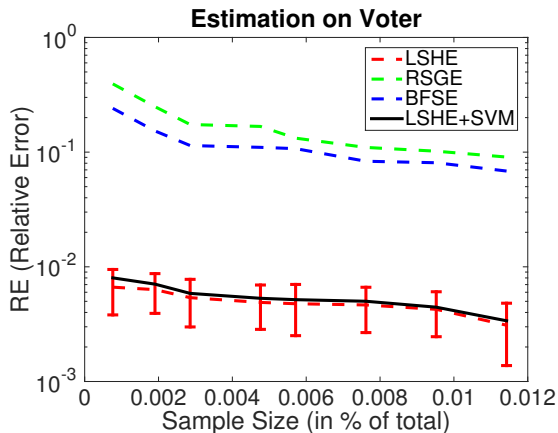
- Our methods: Red (LSHE) and black (LSHE + SVM)
- Comparisons: blue and green (random sampling)

Syrian data set

- 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!