Module 6: Fellegi-Sunter Method Applied to RLdata500

Rebecca C. Steorts

joint with Olivier Binette

Agenda

Load packages

```
## Loading required package: DBI

## Loading required package: RSQLite

## Loading required package: ff

## Loading required package: bit

##
## Attaching package: 'bit'

## The following object is masked from 'package:base':
##
##     xor

## Attaching package ff

## - getOption("fftempdir")=="/var/folders/bv/xhclmwh90zg08

## - getOption("ffextension")=="ff"

## - getOption("ffdrop")==TRUE
```

RLdata500

```
data(RLdata500)
head(RLdata500)
```

```
##   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
```

```
head(identity.RLdata500)
```

```
## [1]  34  51 115 189  72 142
```

Comparison Vectors

Why do we build record pairs of comparison vectors?

Answer: Reduce the total number of record comparisons.

Comparison Vectors

How do we build record pairs of comparison vectors?

Answer: Use the `compare.dedup` function.

Comparison Vectors

```
# create comparison vectors
rpairs <- compare.dedup(RLdata500,
                        identity = identity.RLdata500)
```

Comparison Vectors

```
# inspect comparison vectors
rpairs$pairs[1:5,]
```

```
##   id1 id2 fname_c1 fname_c2 lname_c1 lname_c2 by bm bd is_match
## 1   1   2        0       NA        0       NA  0  1  0        0
## 2   1   3        0       NA        0       NA  0  0  0        0
## 3   1   4        0       NA        0       NA  0  0  0        0
## 4   1   5        0       NA        0       NA  0  0  0        0
## 5   1   6        0       NA        0       NA  0  1  0        0
```

Blocking

Blocking is the reduction of the amount of data pairs through focusing on specified agreement patterns.

Blocking is a common strategy to reduce computation time and memory consumption by only comparing records with equal values for a subset of attributes, called blocking fields.

Blocking

A blocking specification can be supplied to the `compare` function via the argument `blockfld`.

We will consider a blocking pattern where two records must agree in either the **first component of the first name** or **full date of birth**.

Blocking and Comparison Vectors

```
# blocking and comparison vectors
rpairs <- compare.dedup(RLdata500,
                        blockfld = list(1,5:7),
                        identity = identity.RLdata500)
```

Blocking and Comparison Vectors

```
# inspect comparison vectors
rpairs$pairs[c(1:3, 1203:1204),]
```

```
##       id1 id2 fname_c1 fname_c2 lname_c1 lname_c2 by bm bd is_match
## 1       1 174         1       NA         0       NA  0  0  0        0
## 2       1 204         1       NA         0       NA  0  0  0        0
## 3       2   7         1       NA         0       NA  0  0  0        0
## 1203 448 497         1       NA         0       NA  0  0  0        0
## 1204 450 477         1       NA         0       NA  0  0  0        0
```