Module X: fastLink with blocking

Rebecca C. Steorts

Reading

- ▶ Binette and Steorts (2020)
- ► Edmorando et al. (2020)
- ► Fellegi and Sunter (1969)

Agenda

- We continue with our exploration of fastLink by adding blocking.
- ▶ We investigate this using the RLdata10000 data set

Load Packages

```
# load libraries
library(fastLink)
```

Load RLdata10000

```
# 1.0a.d. RI.da.ta.10000
records <-
 read.table("data/RLdata10000.csv",
         sep=",", header=TRUE)
head(records, 4)
##
   FRANK
             <NA> MUELLER
## 1
                            <NA> 1967 9 27
                                             1
                                                3606
     MARTIN
           <NA> SCHWARZ <NA> 1967 2 17
                                                2560
## 2
## 3
    HERBERT <NA> ZIMMERMANN <NA> 1961 11 6
                                                3892
                                                329
## 4
      HANS
           <NA>
                   SCHMITT <NA> 1945 8 14
```

RLdata10000

```
# Number of unique records
length(unique(records$ent_id))
```

Linkage Fields

```
# linkage fields
linkageFields <- c("fname_c1", "lname_c1", "by", "bm", "bd")</pre>
```

Add Numberic Fields

```
# We can add numeric comparisons using dissimilarity
numericMatchFields <- c("by")

# Make sure these are class numeric
records$by <- as.numeric(records$by)</pre>
```

Preparation

```
# linkage fields
linkageFields <- c("fname_c1",</pre>
                    "lname_c1", "by", "bm", "bd")
# string distance fields
stringDistFields <- c("fname_c1", "lname_c1")</pre>
# partial distance fields (fields where we allow
# for agree, disagree, and partially agree)
partialMatchFields <- c("fname c1", "lname c1")
```

Run fastLink

##

==========

##

Setting return.all to FALSE.

How did we do?

Blocking

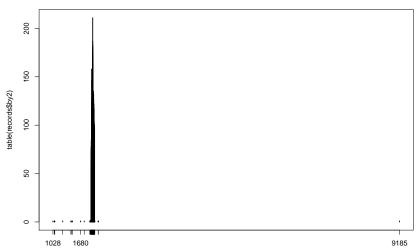
- 1. We can use traditional, deterministic blocking, which is simple and easier but does not allow us to propagate error
- 2. We can also use probabilistic blocking, which allows us propagate (some) of the record linkage error

We will create a simplistic blocking scheme based upon date of birth year.

records\$by2 <- records\$by</pre>

Plot

plot(table(records\$by2))



We will filter out records that are not typical for date of birth year for computational reasons.

```
head(records$by2[records$by < 1924] <- 1923)

## [1] 1923

head(records$by2[records$by > 2008] <- 2009)

## [1] 2009
```

```
blockby <- blockData(records, records, varnames = "by2")
##
## ==========
## blockData(): Blocking Methods for Record Linkage
  ===============
##
  Blocking variables.
##
      Blocking variable by 2 using exact blocking.
##
  Combining blocked variables for final blocking assignment
```

```
# modify the list of linkage fields
# birth year is of no use for
# merging within a traditional block
linkageFields2 <- c("fname_c1", "lname_c1", "bm", "bd")</pre>
```

```
Traditional Blocking
    # store the results from each block
    results <- list()
    for(j in 1:length(blockby)) {
      # subset original data to form block
      records.temp <- records[blockby[[j]]$dfA.inds, ]
      # fastLink applied to a block
      out.temp <- fastLink(dfA = records.temp, dfB =</pre>
                  records.temp,
                  varnames = linkageFields2,
                  stringdist.match = stringDistFields,
                  partial.match = partialMatchFields,
                  cut.a = 0.92, cut.p = 0.84,
                  threshold.match = 0.90,
                  dedupe = FALSE)
      # get the data
      records.temp <-
        getMatches(dfA = records.temp,
                   dfB = records.temp,
                   fl.out = out.temp)
        # adjust the unique identifier to be block specific
      records.temp$dedupe.ids <- paste0("B", j, "_", records.temp$dedupe.id
      # Store the deduplicated data in our object for storage
      results[[j]] <- records.temp
                                                                        18 / 20
```

How many unique entities?

[1] 9145

```
# aggregate all the results from the traditional blocking
recordsfL.blockE <- do.call('rbind', results)
# unique records under fastLink with blocking
length(unique(recordsfL.blockE$dedupe.ids))</pre>
```

Setup

```
library(data.table)
trueMembership <- records$ent_id</pre>
recordIds <- records$rec_id
numRecords <- dim(records)[1]</pre>
matches <-
  data.table(cbind(results$matches$inds.a,
                    results\smatches\sinds.b))
head(matches)
## Null data.table (0 rows and 0 cols)
dim(matches)[1]
```