Introduction to blocking techniques

Brenda Betancourt

Duke University
Department of Statistical Science
bb222@stat.duke.edu

February 8, 2018

Motivation

- Naively matching two files or finding duplicates within a file requires comparing all pairs of records.
- Infeasible for large files even when the comparisons are computationally inexpensive.
- The number of record pairs grows quadratically with the size of the dataset
  - Just 5,000 records $\rightarrow$ 12,497,500 comparisons!

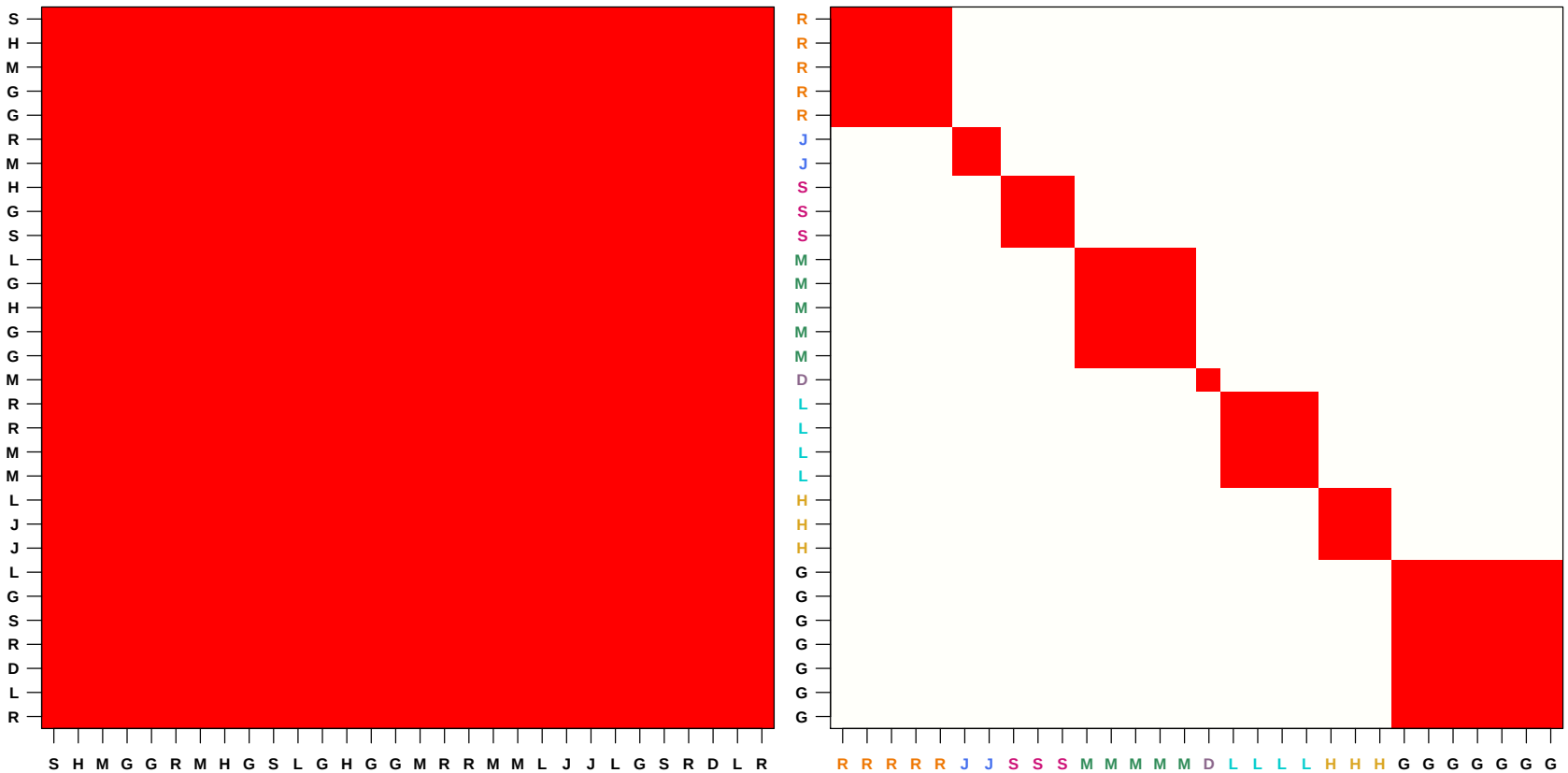
What is blocking?

Technique to reduce the comparison space:

- Filter out dissimilar record pairs that are extremely unlikely to be matches.
  - Perform record linkage only within blocks
- Traditional blocking : compare record pairs that match on one or more keys.
- Cover true matches by using reliable features in the data.

Example: Traditional blocking

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



Example: RLdata500

```
library(RecordLinkage)
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
```

Continuation: RLdata500

```
# Record pairs for comparison
choose(500,2)
```

```
## [1] 124750
```

```
# Blocking by last name initial
last_init <- substr(RLdata500[,"lname_c1"], 1, 1)
head(last_init)
```

```
## [1] "M" "B" "H" "W" "K" "F"
```

```
# Number of blocks
length(unique(last_init))
```

```
## [1] 20
```

Continuation: RLdata500

```
# Number of records per block
tbl <- table(last_init)
head(tbl)
```

```
## last_init
##  A  B  D  E  F  G
##  5 56  2  6 38 12
```

```
# Block sizes can vary a lot
summary(as.numeric(tbl))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    5.75    8.00   25.00   40.00  115.00
```

Continuation: RLdata500

```
# Number of records pairs per block
sapply(tbl, choose, k=2)
```

```
##     A     B     D     E     F     G     H     J     K     L     M
##    10 1540     1    15   703    66   496    28 1035    78 2850
##     S     T     V     W     Z
## 6555     1    21 1326    10
```

```
# Reduction on comparison space
sum(sapply(tbl, choose, k=2))
```

```
## [1] 14805
```

How to choose the blocking key or keys

- Relatively noise free fields in the data.
- More complex blocking schemes can be constructed using disjunctions of conjunctions.
  - Retain only pairs which agree on either lastname initial or/and gender

Example: Voter Survey data

Add data description and ask Which fields are reliable for blocking in this example?

Blocking caveats

- Missing true matches
- Tradeoff between block sizes for computational efficiency (running processes in parallel) and increased false negative rates

Fuzzy Blocking

- Talk about bigram indexing and maybe canopy clustering