UNIVERSITY COLLEGE LONDON

A Short Demo for the Anonymization Procedure

Sinan Shi

October 26, 2016

```
library(ccdata)
library(ccanonym)
```

The YAML configuration file

Create a YAML configuration file as such, where the **identifiable variables** (dirctVars), **key categorical variables** (keyVars), **key numerical variables** (numVars), **key date-time variables** (datetimeVars), **sensitive variables** (sensVars) and the corresponding operations and thresholds are specified.

```
directVars:
    - pasno     # PAS number
    - ICNNO     # Site code
    - ADNO      # INCNARC admission number
    - NHSNO     # NHS number
    - TUADNO    # Transferrring unit admission number
    - DOB           # Date of birth

keyVars:
    - GPCODE    # GP code
    - SEX           # Sex
    - PCODE     # Postcode

sensVars:
    - BPC       # Biopsy proven cirrhosis
    - AIDS_V3 # HIC/AIDS
    - PH            # Portal hypertension
    - RAICU1    # Primary reason for admission to your unit
    - RAICU2    # Secondary reasons for admission to your unit
    - URAICU    # Ultimate primary reason for admission to unit

numVars:
    HCM: # Height
        microaggregation:
            aggr: 2

datetimeVars:
    DOAH:  # Date of original admission to/attendance at acute hospital'
        microaggregation:
```

```
            aggr: 1
        addNoise:
            noise: 2

    DAH: # Date of admission to your hospital
        addNoise:
            noise: 2

    DOAICU:  # Date of original admission to ICU/HDU'
        microaggregation:
            aggr: 1
        addNoise:
            noise: 1
... ...
```

```
conf <- yaml.load_file("../data/test_demo.yaml")
vars <- anony.var(conf)
```

Identifiable data set

The identifiable data set is usually stored in ccRecord format. In the following code, we create the ccRecord object from a XML file which contains only five episodes.

```
ccd <- xml2Data("../tests/data/test_data_anonym.xml")
demg.table <- as.data.frame(sql.demographic.table(ccd))
```

```
all.var <- c(vars$dirv, vars$all.vars) # all variables besides non-confidential data
pander(demg.table[, all.var],  style = 'rmarkdown')
```

Table 0.1: Table continues below

pasno	ICNNO	NHSNO	DOB	GPCODE	SEX	PCODE	HCM	DAH	DUDICU
pas_1	site_1	nhs_1	1988-06-07	GPCODE1	F	NW1 1BB	170	2014-02-01	NULL
pas_2	site_1	nhs_2	1980-12-30	GPCODE1	F	NW1 1BB	174	2014-02-01	NULL
pas_1	site_1	nhs_1	1970-01-25	GPCODE2	F	NW1 1BB	170	2014-02-09	NULL
pas_1	site_1	nhs_1	1977-01-25	GPCODE2	M	NW1 2BB	170	2014-02-01	NULL
pas_1	site_1	nhs_1	1955-01-04	GPCODE4	M	NULL	170	2014-02-01	NULL

DDBSD	DWFRD	TWFRD	DDH	DAICU	DDICU	AIDS_V3
NULL	2014-02-01	18:00:00	2014-02-01	2014-02-01 10:00	2014-02-10	TRUE
NULL	2014-02-01	10:00:00	2014-02-01	2014-02-01 10:00	NULL	FALSE
NULL	2014-02-08	18:00:00	2014-02-26	2014-02-01 10:00	2014-02-10	FALSE
NULL	2014-02-01	18:00:00	2014-02-27	2014-02-01 10:00	2014-02-10	FALSE
NULL	2014-02-01	18:00:00	2014-02-01	2014-02-01 10:00	2014-02-10	FALSE

Anonymisation

```
anonccd <- anonymisation(ccd, conf="../data/test_demo.yaml")#, remove.alive=T, verbose=T)
pander(sql.demographic.table(anonccd)[, all.var, with=F], style="rmarkdown")
```

Table 0.3: Table continues below

pasno	ICNNO	NHSNO	DOB	GPCODE	SEX	PCODE	HCM	DAH
NULL	NULL	NULL	NULL	GPCODE1	F	NW1 1BB	170	2014-02-01 02:15:39
NULL	NULL	NULL	NULL	GPCODE1	F	NW1 1BB	171.3	2014-02-01 00:25:32
NULL	NULL	NULL	NULL	NULL	F	NW1 1BB	170	2014-02-09 01:23:00
NULL	NULL	NULL	NULL	NULL	M	NW1 2BB	171.3	2014-02-01 00:36:22
NULL	NULL	NULL	NULL	GPCODE4	M	NULL	171.3	2014-02-01 01:25:51

Table 0.4: Table continues below

DUDICU	DDBSD	DWFRD	TWFRD	DDH	DAICU	DDICU
NULL	NULL	2014-01-31 23:25:46	18:00	2014-02-01 01:56:21	2014-02-01 10:00	NULL
NULL	NULL	2014-02-01 01:49:28	10:02	2014-01-31 19:14:47	2014-02-01 10:00	NULL
NULL	NULL	2014-02-08 00:26:20	18:01	2014-02-26 05:50:48	2014-02-01 10:00	NULL
NULL	NULL	2014-01-31 23:34:47	18:00	2014-02-26 18:30:40	2014-02-01 10:00	NULL
NULL	NULL	2014-02-01 00:14:17	18:01	2014-02-01 05:38:00	2014-02-01 10:00	NULL

AIDS_V3
TRUE
FALSE
FALSE
FALSE
FALSE