Validação

José Elvano Moraes

4/15/2021

```
## 
## -- Column specification ---------------------------------------------------------
## cols(
##   .default = col_double(),
##   DT_NOTIFIC = col_character(),
##   DT_SIN_PRI = col_character(),
##   SG_UF_NOT = col_character(),
##   ID_REGIONA = col_character(),
##   ID_MUNICIP = col_character(),
##   ID_UNIDADE = col_character(),
##   CS_SEXO = col_character(),
##   DT_NASC = col_character(),
##   CS_ETINIA = col_logical(),
##   ID_PAIS = col_character(),
##   SG_UF = col_character(),
##   ID_RG_RESI = col_character(),
##   ID_MN_RESI = col_character(),
##   OUTRO_DES = col_character(),
##   FATOR_RISC = col_character(),
##   OBES_IMC = col_character(),
##   MORB_DESC = col_character(),
##   DT_UT_DOSE = col_character(),
##   DT_VAC_MAE = col_logical(),
##   DT_DOSEUNI = col_logical()
##   # ... with 54 more columns
## )
## i Use `spec()` for the full column specifications.

## Warning: 1221 parsing failures.
##  row       col           expected          actual                          file
## 1857 OUT_ANIM  1/0/T/F/TRUE/FALSE CACHORRO       'dados/INFLUD21-05-04-2021 copy.csv'
## 2266 CS_ETINIA 1/0/T/F/TRUE/FALSE ANAMBE        'dados/INFLUD21-05-04-2021 copy.csv'
## 2352 DS_AN_OUT 1/0/T/F/TRUE/FALSE RHINOVIRUS     'dados/INFLUD21-05-04-2021 copy.csv'
## 3352 OUT_ANIM  1/0/T/F/TRUE/FALSE CACHORRO       'dados/INFLUD21-05-04-2021 copy.csv'
## 3861 OUT_ANIM  1/0/T/F/TRUE/FALSE GATO, CACHORRO 'dados/INFLUD21-05-04-2021 copy.csv'
## .... ......... .................. .............. ....................................
## See problems(...) for more details.

## Rows: 76,666
## Columns: 18
## $ IDADE     <dbl> 72, 57, 79, 55, 81, 46, 57, 89, 49, 66, 58, 86, 73, 75, 49,~
## $ FEBRE     <fct> 2, 2, 2, 1, 2, 1, 1, 1, 1, 2, 1, 2, 1, 2, 2, 1, 1, 2, 2, 2,~
## $ GARGANTA  <fct> 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 2, 2, 2, 2,~
## $ DISPNEIA  <fct> 1, 2, 2, 1, 2, 2, 2, 2, 1, 1, 1, 1, 2, 1, 2, 1, 2, 1, 1, 2,~
```

```
## $ SATURACAO  <fct> 1, 1, 2, 2, 2, 2, 2, 1, 2, 2, 1, 1, 1, 1, 2, 2, 2, 1, 2, 2,~
## $ EVOLUCAO   <fct> 2, 1, 1, 1, 2, 1, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 1, 1, 1, 1,~
## $ RENAL      <fct> 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
## $ DIABETES   <fct> 2, 1, 2, 1, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2,~
## $ OBESIDADE  <fct> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 1, 2, 2, 9, 2,~
## $ PERD_OLFT  <fct> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2,~
## $ PERD_PALA  <fct> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
## $ NEUROLOGIC <fct> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 9, 2, 2, 2,~
## $ PNEUMOPATI <fct> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
## $ UTI        <fct> 1, 1, 2, 1, 2, 2, 1, 2, 2, 2, 2, 2, 1, 1, 2, 2, 1, 2, 2, 2,~
## $ CARDIOPATI <fct> 2, 2, 2, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1,~
## $ SUPORT_VEN <fct> 2, 2, 2, 9, 3, 3, 2, 2, 3, 2, 2, 3, 1, 1, 3, 3, 2, 3, 9, 3,~
## $ ASMA       <fct> 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
## $ ANTIVIRAL  <fct> 2, 2, 2, 9, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~

## Rows: 76,666
## Columns: 18
## $ IDADE      <fct> "(37,73]", "(37,73]", "(73,109]", "(37,73]", "(73,109]", "(~
## $ FEBRE      <fct> 2, 2, 2, 1, 2, 1, 1, 1, 1, 2, 1, 2, 1, 2, 2, 1, 1, 2, 2, 2,~
## $ GARGANTA   <fct> 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 2, 2, 2, 2,~
## $ DISPNEIA   <fct> 1, 2, 2, 1, 2, 2, 2, 2, 1, 1, 1, 1, 2, 1, 2, 1, 2, 1, 1, 2,~
## $ SATURACAO  <fct> 1, 1, 2, 2, 2, 2, 2, 1, 2, 2, 1, 1, 1, 1, 2, 2, 2, 1, 2, 2,~
## $ EVOLUCAO   <fct> 2, 1, 1, 1, 2, 1, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 1, 1, 1, 1,~
## $ RENAL      <fct> 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
## $ DIABETES   <fct> 2, 1, 2, 1, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2,~
## $ OBESIDADE  <fct> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 1, 2, 2, 9, 2,~
## $ PERD_OLFT  <fct> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2,~
## $ PERD_PALA  <fct> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
## $ NEUROLOGIC <fct> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 9, 2, 2, 2,~
## $ PNEUMOPATI <fct> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
## $ UTI        <fct> 1, 1, 2, 1, 2, 2, 1, 2, 2, 2, 2, 2, 1, 1, 2, 2, 1, 2, 2, 2,~
## $ CARDIOPATI <fct> 2, 2, 2, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1,~
## $ SUPORT_VEN <fct> 2, 2, 2, 9, 3, 3, 2, 2, 3, 2, 2, 3, 1, 1, 3, 3, 2, 3, 9, 3,~
## $ ASMA       <fct> 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
## $ ANTIVIRAL  <fct> 2, 2, 2, 9, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~

## White list:

##        from        to
## [1,] "OBESIDADE" "DIABETES"
## [2,] "IDADE"     "DIABETES"
## [3,] "SATURACAO" "SUPORT_VEN"
## [4,] "IDADE"     "SUPORT_VEN"
## [5,] "DIABETES"  "RENAL"
## [6,] "DIABETES"  "CARDIOPATI"
## [7,] "UTI"       "EVOLUCAO"
## [8,] "RENAL"     "CARDIOPATI"
```
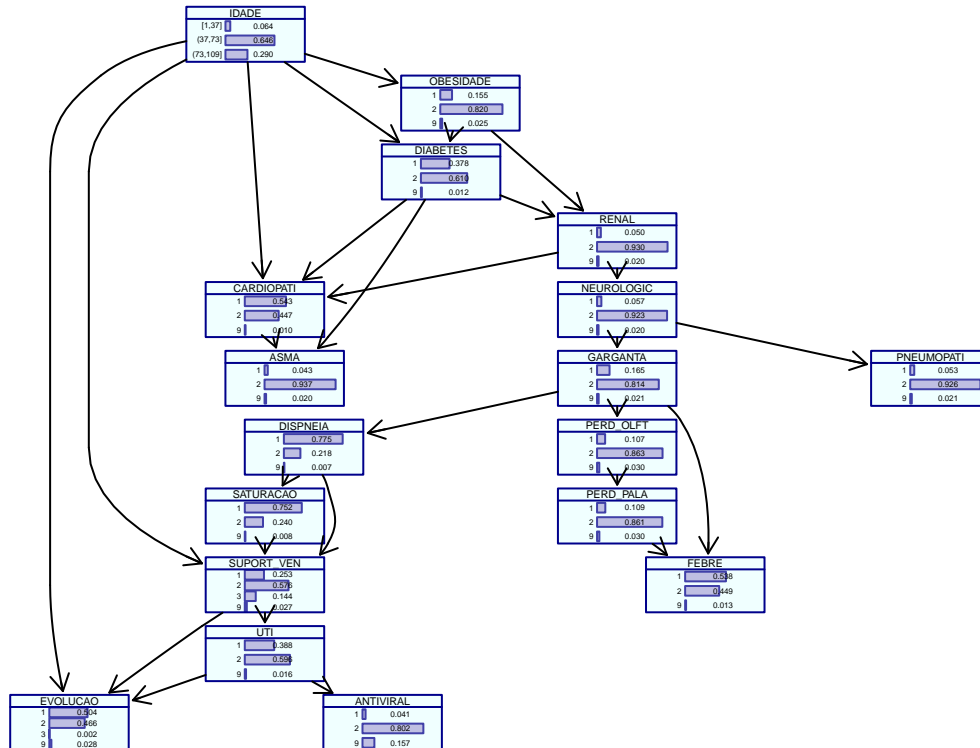
...
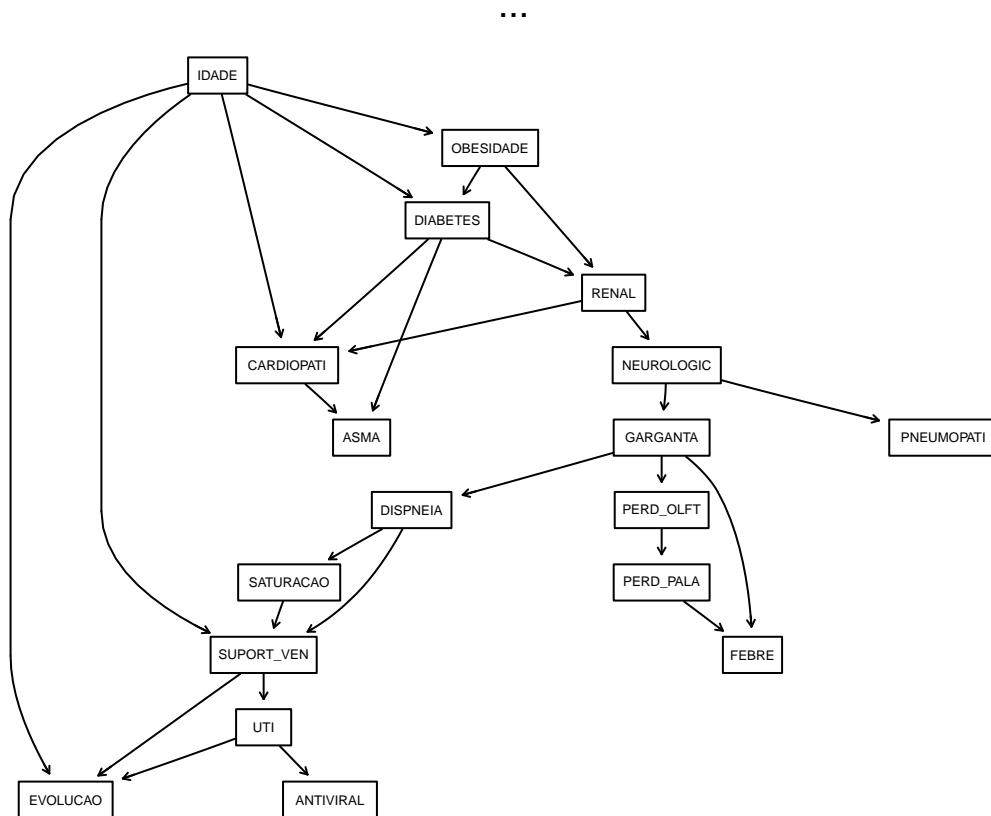

```
## Warning in from.bn.fit.to.grain(x): NaN conditional probabilities in CARDIOPATI,
## replaced with a uniform distribution.
```

Rede de probabilidades

IDADE
[1,37] 0.064
(37,73) 0.646
(73,109] 0.290

OBESIDADE
1 0.155
2 0.820
9 0.025

DIABETES
1 0.378
2 0.610
9 0.012

RENAL
1 0.050
2 0.930
9 0.020

CARDIOPATI
1 0.543
2 0.447
9 0.010

NEUROLOGIC
1 0.057
2 0.923
9 0.020

PNEUMOPATI
1 0.053
2 0.926
9 0.021

ASMA
1 0.043
2 0.937
9 0.020

GARGANTA
1 0.165
2 0.814
9 0.021

DISPNEIA
1 0.775
2 0.218
9 0.007

PERD_OLFT
1 0.107
2 0.863
9 0.030

SATURACAO
1 0.752
2 0.240
9 0.008

PERD_PALA
1 0.109
2 0.861
9 0.030

SUPORT_VEN
1 0.253
2 0.576
3 0.144
9 0.027

FEBRE
1 0.538
2 0.449
9 0.013

UTI
1 0.388
2 0.596
9 0.016

EVOLUCAO
1 0.504
2 0.466
3 0.002
9 0.028

ANTIVIRAL
1 0.041
2 0.802
9 0.157

Bootstraping

```r
boots.trap  <- 200
str.diff = suppressMessages(boot.strength(s1, R = boots.trap, algorithm = "mmhc"))
cat(paste('Threshold: ', attr(str.diff, "threshold")))
```

```
## Threshold:  0.5
```

```r
avg.diff = averaged.network(str.diff)
strength.plot(avg.diff, str.diff, shape = "rectangle", main = paste("Iterações = ", boots.trap))
```

Iterações = 200
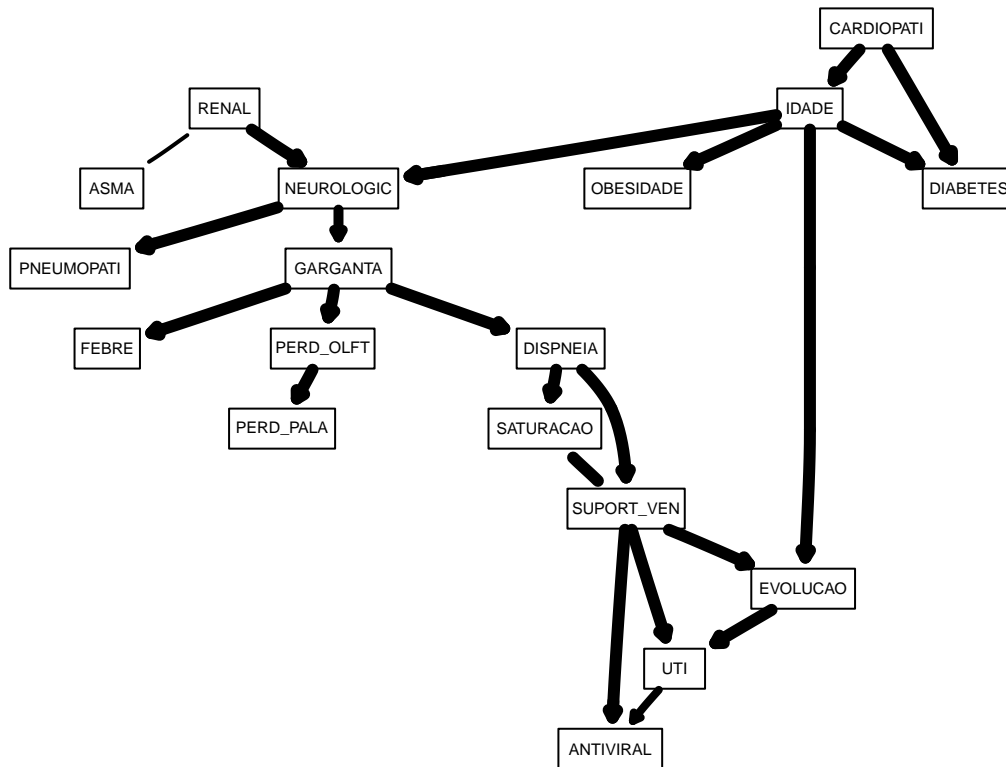
How can we compare the averaged network (avg.diff) with the network we originally learned in from all the data? The most qualitative way is to plot the two networks side by side, with the nodes in the same positions, and highlight the arcs that appear in one network and not in the other, or that appear with different directions.
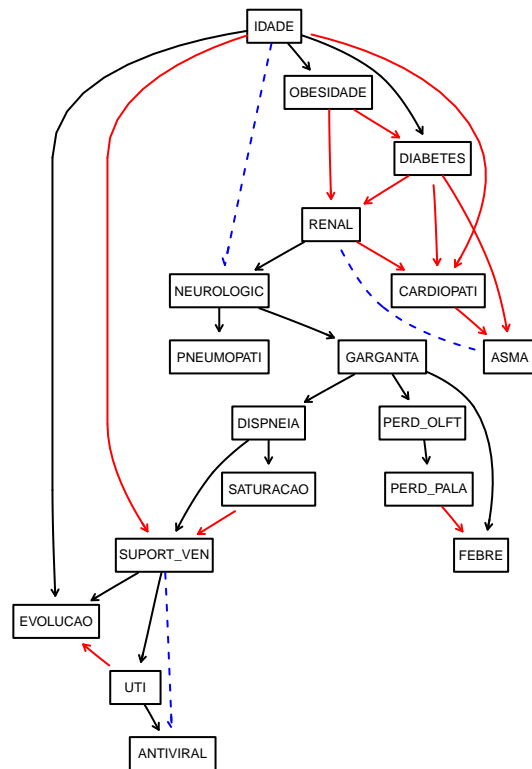
```
par(mfrow = c(1, 2))
graphviz.compare(avg.diff, bn1, shape = "rectangle", main = c("DAG médio", "DAG único"))
```
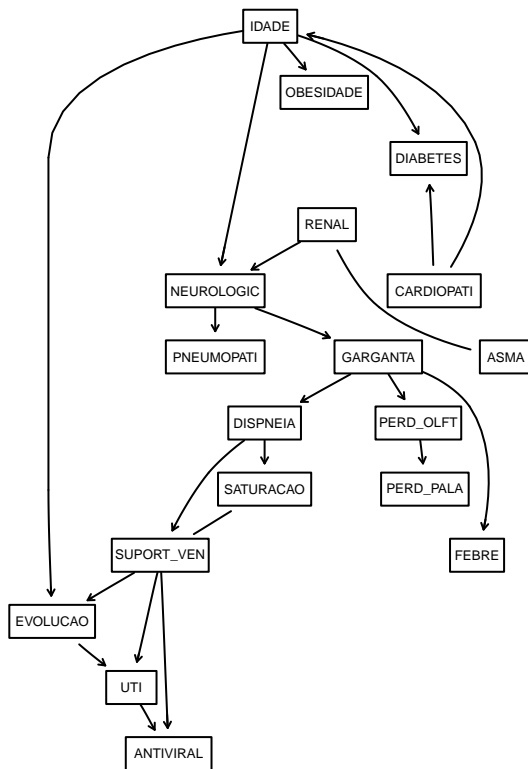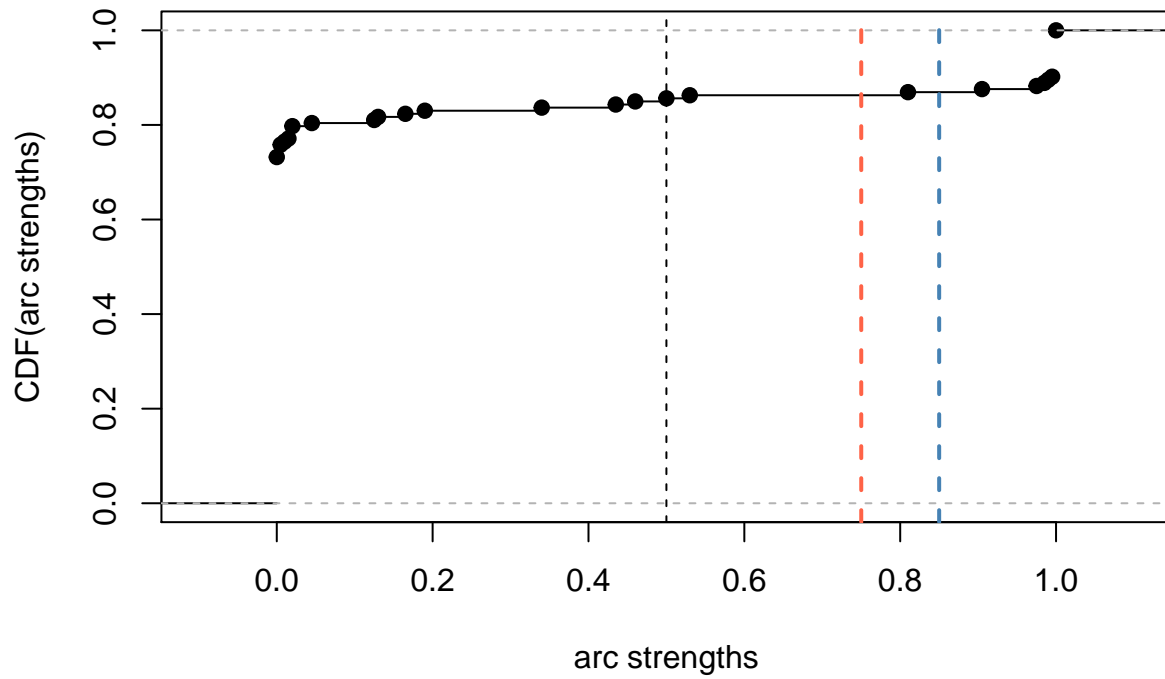
DAG médio   DAG único

It is also a good idea to look at the threshold with respect to the distribution of the arc strengths

```
plot(str.diff)
abline(v = 0.75, col = "tomato", lty = 2, lwd = 2)
abline(v = 0.85, col = "steelblue", lty = 2, lwd = 2)
```
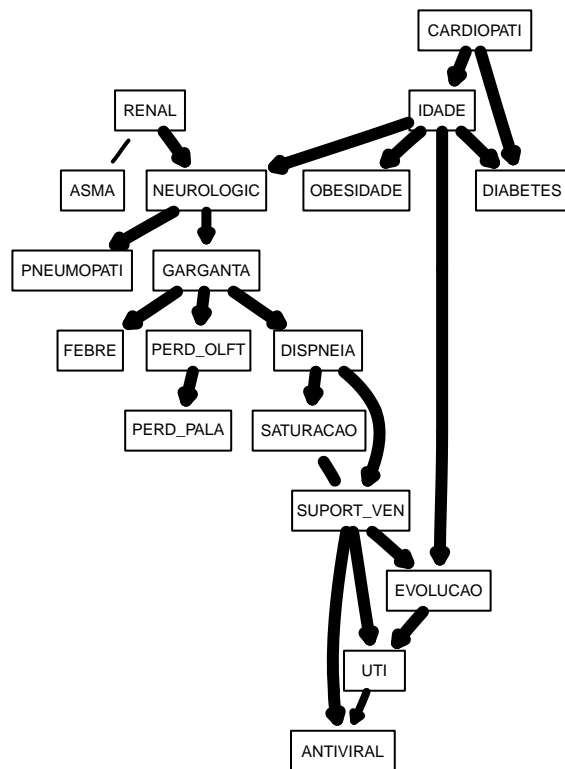
threshold = 0.5

The simpler network we obtain by setting threshold = 0.8 in averaged.network() is shown below; it is certainly easier to reason with from a qualitative point of view.

```
par(mfrow = c(1, 2))
avg.simpler = averaged.network(str.diff, threshold = 0.75)
strength.plot(avg.diff, str.diff, shape = "rectangle", main = paste("Iterações = ", boots.trap, " Thr =
strength.plot(avg.simpler, str.diff, shape = "rectangle", main = 'Iterações = 100 Thr = 0.75')
```

Iterações = 200 Thr = 0.5

Iterações = 100 Thr = 0.75