stat992HW1

SongWang

10/08/2015

read the payment data and select a subset from it

conditional on the state = CA. select subset of NIPs where state = CA, entity_type = "individual"

```
rm(list=ls())
library(data.table)  # so fast!
# install.packages('igraph')
library(igraph)  # all the basic graph operations.
```

```
##
## Attaching package: 'igraph'
##
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
##
## The following object is masked from 'package:base':
##
##     union
```

```
###############
setwd("~/Stat/Courses/Physisian_Referral_Network/RScripts/")
DataPath <- "../Data/"
ResultsPath <- "../Results/"
PlotsPath <- "../Plots/"


#### payment data

# Payment = fread(paste0(DataPath,
#       "Medicare_Provider_Util_Payment_PUF_CY2013/Medicare_Provider_Util_Payment_PUF_CY2013.txt"),
#                 sep = "\t")
# Payment <- Payment[-1]
# setkey(Payment, NPI)
# head(Payment)
#
# Payment_NPI_ca <- Payment[NPPES_PROVIDER_STATE=="CA"&NPPES_ENTITY_CODE=="I"]
# Payment_NPI_total_ca= Payment_NPI_ca[,.(NPI,totalPay=AVERAGE_MEDICARE_ALLOWED_AMT * LINE_SRVC_CNT)]
# Payment_NPI_total_ca <- Payment_NPI_total_ca[,.(totalPay=sum(totalPay)),by=NPI]
#
# save(Payment_NPI_total_ca,file = paste0(DataPath, "Payment_NPI_total_ca.RData"))


system.time(load(paste0(DataPath, "EtDT.RData")))
```

```
##    user  system elapsed
##  63.217   0.643  64.127
```

```
system.time(load(paste0(DataPath,"Payment_NPI_total_ca.RData")))
```

```
##    user  system elapsed
##   0.006   0.000   0.007
```

```
## Payment_NPI_total_ca
NPI_SF <- DT[City=="SAN FRANCISCO" & NPI%in%Payment_NPI_total_ca$NPI ]  ## physisian --individual in ca
setkey(NPI_SF,NPI)
#NPI_SF = NPI_SF[unique(NPI_SF$NPI), mult="first"]
Edge_SF <- Et[V1 %in% unique(NPI_SF$NPI)]
setkey(Edge_SF, V1)

setkey(Payment_NPI_total_ca,NPI)
Payment_SF <- Payment_NPI_total_ca[NPI%in%NPI_SF$NPI]
Payment_SF<- Payment_SF[,.(NPI,totalPay,logPay = log(totalPay+1))]

paylevel <- function(x){
    high <- quantile(x,probs = 0.90)
    high_medium <- quantile(x,probs = 0.70)
    low_medium <- quantile(x,probs = 0.30)
    low <- quantile(x,probs = 0.10)
    y <- as.character(x)
    y[which(x>=high)]="high"
    y[which(x<high &x>= high_medium)] ="high_medium"
    y[which(x<high_medium &x>= low_medium)] ="medium"
    y[which(x<low_medium &x>= low)] ="low_medium"
    y[which(x<low)] ="low"
    y[is.na(x)]="NA"
    return(y)
}
Payment_SF <-Payment_SF[,.(NPI,totalPay,logPay,payLevel=paylevel(totalPay))]
setkey(Payment_SF,NPI)
```

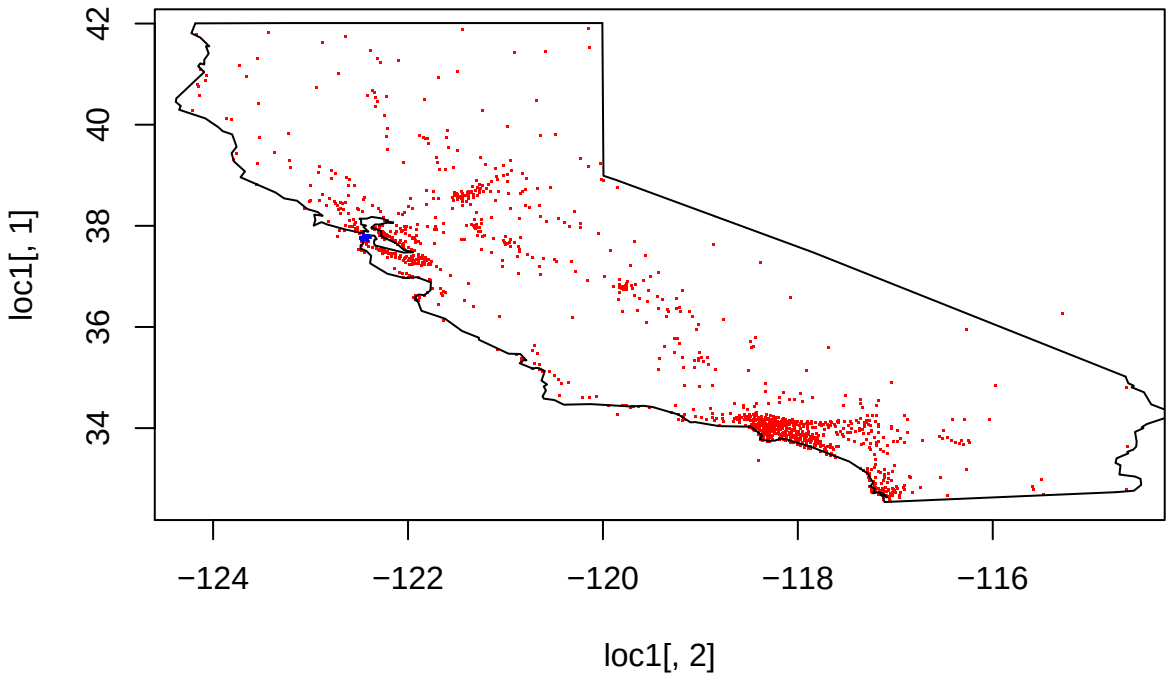
Look at the positions of Physician in San Francisco.

```
library(zipcode)
zip = NPI_SF[as.character(Payment_SF$NPI)]$"Zip Code"
zip = substr(zip, start = 1, stop = 5)

data(zipcode)   # this contains the locations of zip codes
zipcode = as.data.table(zipcode); setkey(zipcode, zip)
loc =  zipcode[zip, c("latitude", "longitude"), with = F]
loc = loc[complete.cases(loc)]
loc = data.frame(loc)



### show the geographic positions
library(maps); library(ggplot2)
```

```
##
```

```
##  # ATTENTION: maps v3.0 has an updated 'world' map.        #
##  # Many country borders and names have changed since 1990. #
##  # Type '?world' or 'news(package="maps")'. See README_v3. #
```

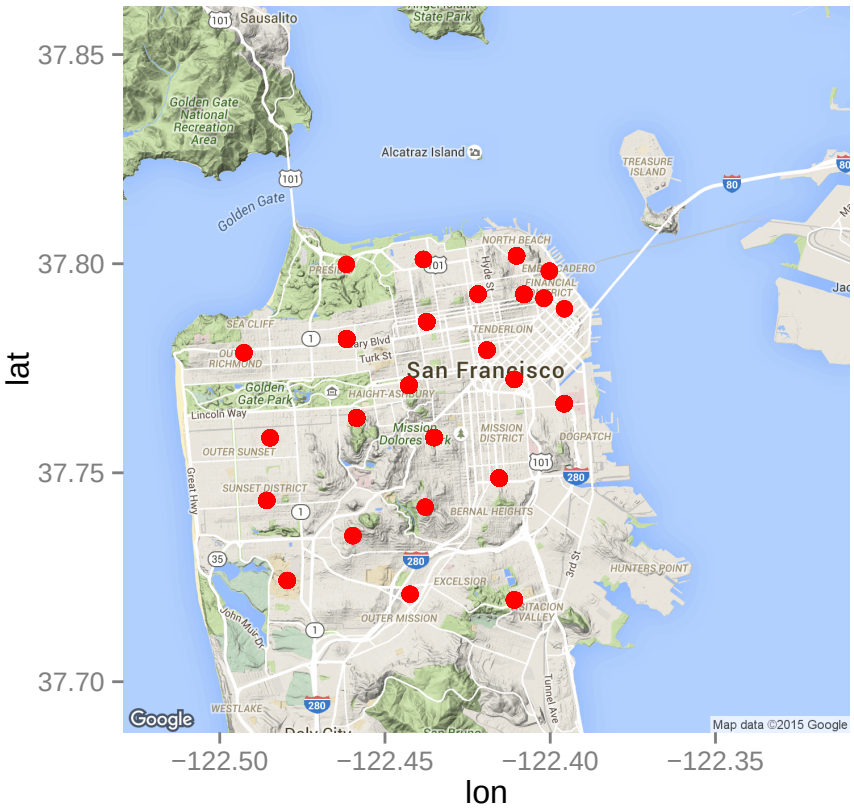
```
library(ggmap)
ca <- DT[State=="CA"]
zip = ca$"Zip Code"
zip = substr(zip, start = 1, stop = 5)

data(zipcode)   # this contains the locations of zip codes
zipcode = as.data.table(zipcode); setkey(zipcode, zip)
loc1 =  zipcode[zip, c("latitude", "longitude"), with = F]
loc1 = loc1[complete.cases(loc1)]
loc1 = data.frame(loc1)
plot(loc1[,2],loc1[,1], pch=".",col="red")
map(database = 'state', region = c('california'),fill=F, add = T)
points(loc[,2],loc[,1],col="blue",pch=".")
```




```
sfMap = get_map(location = 'San Francisco', zoom = 12)
```

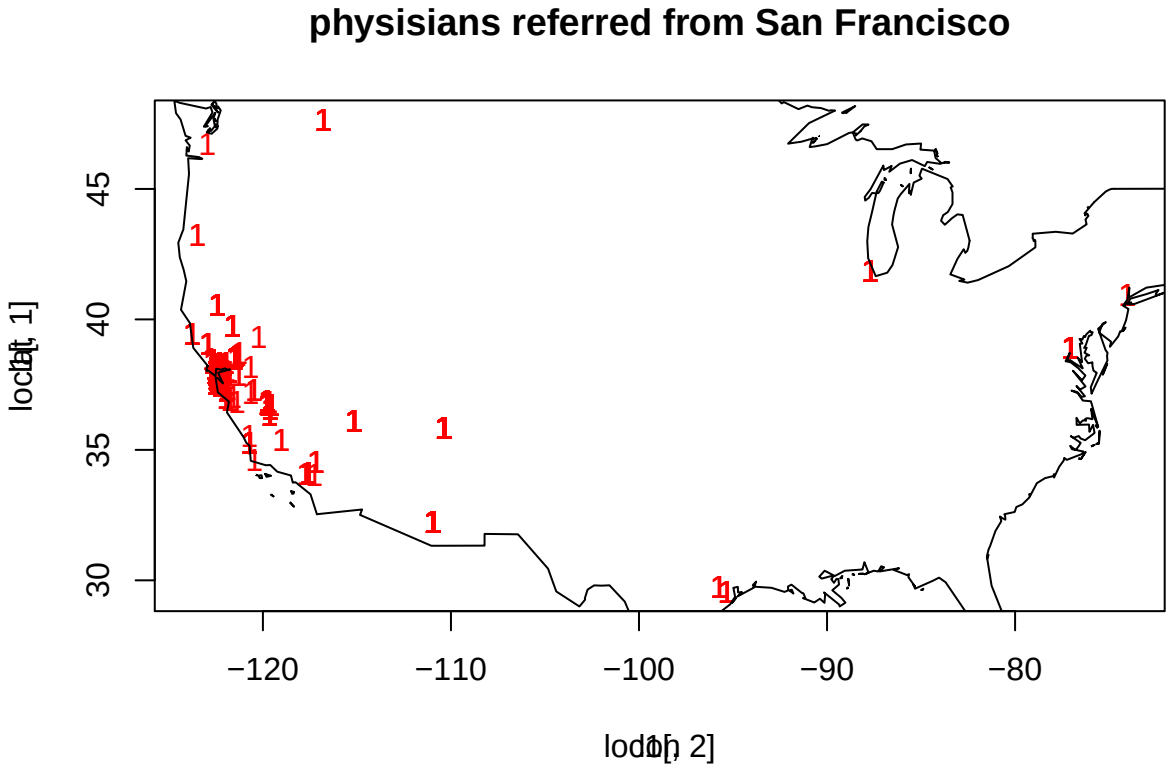
```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=San+Francisco&zoom=12&size=640x6
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=San%20Francisco&sens
```

```
ggmap(sfMap) +  geom_point(data=loc,aes(x = longitude, y = latitude,
                                        position="jitter"),color="red", size=3)
```



take a look at how many physisians are outside the San Francisco. They are located all over the country.

```
Edge_SF <- Edge_SF[V2 %in% V1]

outNode <- Edge_SF[,.(V2)]
zip <- DT[outNode]$"Zip Code"
zip = substr(zip, start = 1, stop = 5)

data(zipcode)    # this contains the locations of zip codes
zipcode = as.data.table(zipcode); setkey(zipcode, zip)
loc1 =  zipcode[zip, c("latitude", "longitude"), with = F]
loc1 = loc1[complete.cases(loc1)]
loc1 = data.frame(loc1)
plot(loc1[,2],loc1[,1], pch="1",col="red")
title(main="physisians referred from San Francisco",font=1, xlab="lon", ylab="lat")
map(database = 'world', region = c('usa'),fill=F, add = T)
```



show the referral network confined to network among physicians in SF, Trying to show the relationship between network and total payment from Medicare

```
Edge_SF1 <- Edge_SF[V2 %in% V1]
el=as.matrix(Edge_SF1)[,1:2] #igraph needs the edgelist to be in matrix format
g=graph.edgelist(el,directed = F) # this creates a graph.
g= simplify(g)  # removes any self loops and multiple edges
vcount(g)
```

```
## [1] 842
```

```
ecount(g)
```

```
## [1] 4544
```

```
cities <- DT[Edge_SF1[,.(V2)]]$City  ## cannot just simply pick one, having multiple address.
sort(table(cities), decreasing=TRUE)[1:30]
```

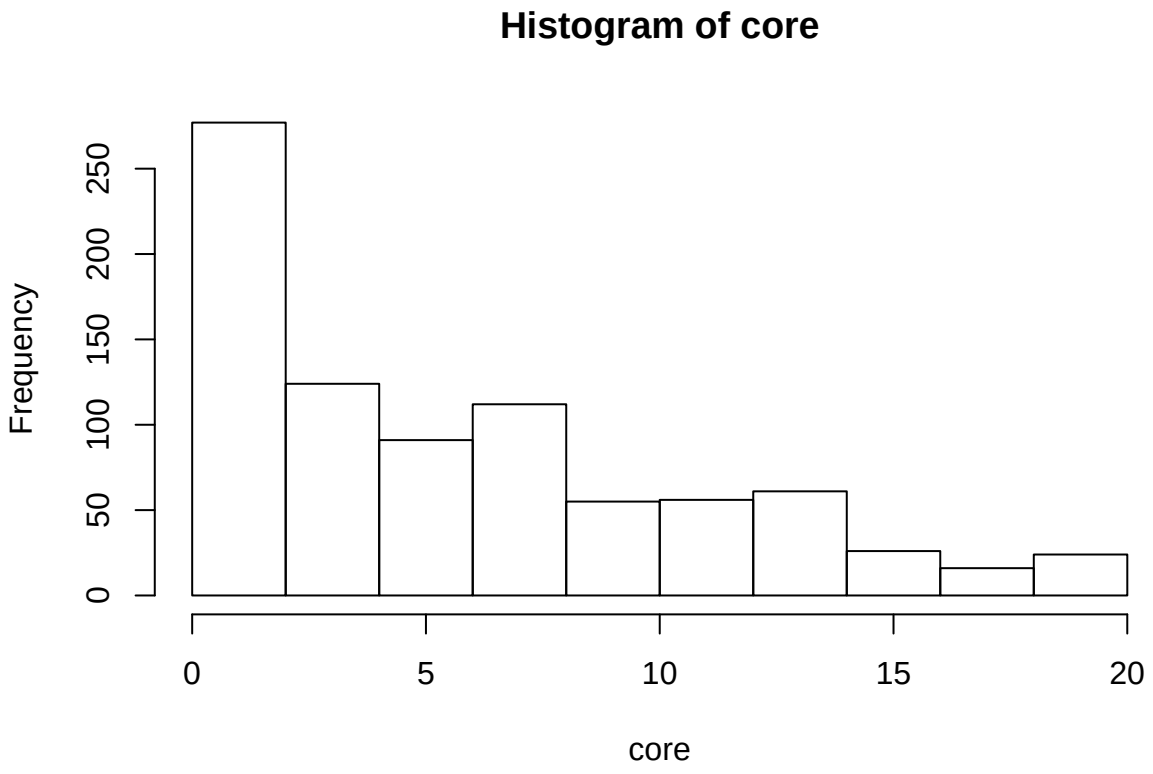
```
## cities
##       SAN FRANCISCO          SANTA ROSA          BURLINGAME
##               21737                 792                 532
##           GREENBRAE           SAN MATEO          SACRAMENTO
##                 433                 422                 253
##        REDWOOD CITY           DALY CITY           KENTFIELD
##                 237                 178                 120
```

```
##              NOVATO              SONOMA              TUCSON
##                 120                 120                 114
##            BERKELEY       WALNUT CREEK            SAN JOSE
##                 104                  68                  63
##          MOSS BEACH            POLACCA           FAIRFIELD
##                  60                  57                  54
##             FRESNO    SAN LUIS OBISPO            PARADISE
##                  45                  36                  30
## SOUTH SAN FRANCISCO           PALO ALTO          WASHINGTON
##                  29                  17                  17
##           LAS VEGAS             REDDING             ANTIOCH
##                  16                  15                  14
##             HANFORD             ORINDA           SAN PABLO
##                  13                  13                  13
```

```
clust <- clusters(g)
clust$csize
```

```
##  [1]   4 745  19  23   5   2   2  17   2   2   2   3   6   2   2   2   2
## [18]   2
```

```
core = graph.coreness(g)  # talk about core.
hist(core)
```

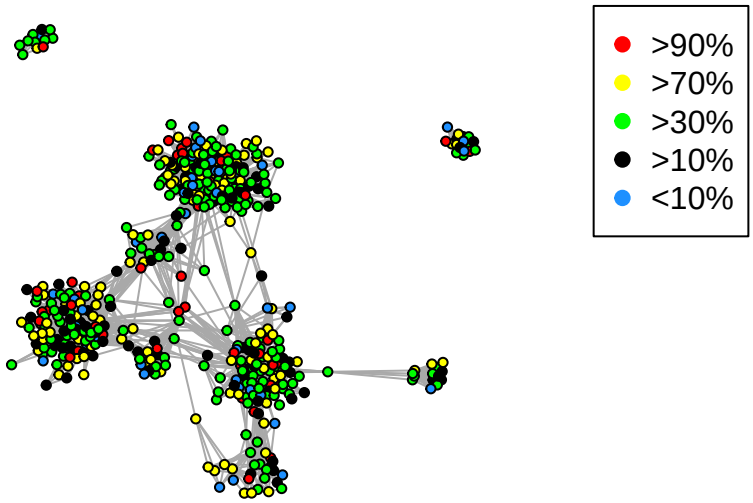



```
sum(core>3)
```

```
## [1] 481
```

```
g1 = induced.subgraph(graph = g,vids = V(g)[core>3])  # talk about induced subgraphs.

layout(1)
v.colors <- as.character(Payment_SF[V(g1)]$payLevel)
v.colors[v.colors=="high"]="red"
v.colors[v.colors=="low"] ="dodgerblue"
v.colors[v.colors=="high_medium"]="yellow"
v.colors[v.colors=="medium"]="green"
v.colors[v.colors=="low_medium"]="black"
set.seed(42)
plot(g1,layout = layout.fruchterman.reingold, vertex.label = NA,
     edge.arrow.size=0.05,  vertex.size=4,
     vertex.color=v.colors)
title(main="total pay for individual physician in San Francisco",cex.main=0.8)
legend("topright",legend=c(">90%",">70%",">30%",">10%","<10%"),
                    col=c("red","yellow","green","black","dodgerblue"),pch=19,
       border = "white")
```

total pay for individual physician in San Francisco



Part 3, Results based on Spectral clustering.