Applied SNA with R

George G. Vega Yon

2021-04-27

Contents

Chapter 1

About this book

This book will be build as part of a workshop on Applied Social Network Analysis with R. Its contents will be populated as the sessions take place, and for now there is particular program that we will follow, instead, we have the following workflow:

1. Participants will share their data and what they need to do with it.
2. Based on their data, I'll be preparing the sessions trying to show attendees how would I approach the problem, and at the same time, teach by example about the R language.
3. Materials will be published on this website and, hopefully, video recordings of the sessions.

At least in the first version, the book will be organized by session, this is, one chapter per session.

All the book materials can be downloaded from `https://github.com/gvegayon/appliedsnar`

In general, we will besides of R itself, we will be using R studio and the following R packages: dplyr for data management, stringr for data cleaning, and of course igraph, netdiffuseR (a bit of a bias here), and statnet for our neat network analysis.[1]

[1]Some of you may be wondering "what about ggplot2 and friends? What about `tidyverse`," well, my short answer is I jumped into R before all of that was that popular. When I started plots were all about `lattice`, and after a couple of years on that, about base R graphics. What I'm saying is that so far I have not find a compelling reason to leave my "old-practices" and embrace all the `tidyverse` movement (religion?).

Chapter 2

Introduction

For this book we need the following

R Core Team (2017b)

1. Install R from CRAN: `https://www.r-project.org/`
2. (optional) Install Rstudio: `https://rstudio.org`

While I find RStudio extreamly useful, it is not necesary to use it with R.

Chapter 3

R Basics

3.1 What is R

A good reference book for both new and advanced user is "The Art of R programming" (Matloff 2011)[1]

3.2 How to install packages

Nowadays there are two ways of installing R packages (that I'm aware of), either using `install.packages`, which is a function shipped with R, or use the devtools R package to install a package from some remote repository other than CRAN, here is a couple of examples:

```
# This will install the igraph package from CRAN
> install.packages("netdiffuseR")

# This will install the bleeding-edge version from the project's github repo!
> devtools::install_github("USCCANA/netdiffuseR")
```

The first one, using `install.packages`, installs the CRAN version of `netdiffuse`R, whereas the second installs whatever version is plublished on `https://github.com/USCCANA/netdiffuseR`, which is usually called the development version.

In some cases users may want/need to install packages from command line as some packages need extra configuration to be installed. But we won't need to look at it now.

[1] Here a free pdf version distributed by the author.

Chapter 4

Network Nomination Data

The data can be downloaded from here.

The codebook for the data provided here is in the appendix.

This chapter's goals are:

1. Read the data into R,
2. Create a network with it,
3. Compute descriptive statistics
4. Visualize the network

4.1 Data preprocessing

4.1.1 Reading the data into R

R has several ways of reading data in. You data can be Raw plain files like CSV, tab delimited or specified by column width, for which you can use the `readr` package (Wickham, Hester, and Francois 2017); or it can be binary files like dta (Stata), Octave, SPSS, for which `foreign` (R Core Team 2017a) can be used; or it could be excel files in which case you should be using `readxl` (Wickham and Bryan 2017). In our case, the data for this session is in Stata format:

```
library(foreign)

# Reading the data
dat <- foreign::read.dta("03-sns.dta")

# Taking a look at the data's first 5 columns and 5 rows
dat[1:5, 1:10]
```

```
##   photoid school hispanic female1 female2 female3 female4 grades1 grades2
```

```
## 1        1    111        1       NA       NA        0        0       NA       NA
## 2        2    111        1        0       NA       NA        0      3.0       NA
## 3        7    111        0        1        1        1        1      5.0      4.5
## 4       13    111        1        1        1        1        1      2.5      2.5
## 5       14    111        1        1        1        1       NA      3.0      3.5
##   grades3
## 1     3.5
## 2      NA
## 3     4.0
## 4     2.5
## 5     3.5
```

4.1.2 Creating a unique id for each participant

Now suppose that we want to create a unique id using the school and photo id. In this case, since both variables are numeric, a good way of doing it is to encode the id such that, for example, the last three x numbers are the photoid and the first ones are the school id. To do this we need to take into account the range of the variables. Here, `photoid` has the following range:

```
(photo_id_ran <- range(dat$photoid))
```

```
## [1]    1 2074
```

As the variable spans up to 2074, we need to set the last 4 units of the variable to store the `photoid`. We will use `dplyr` (Wickham et al. 2017) and `magrittr` (Bache and Wickham 2014)] (the pipe operator, `%>%`) to create this variable, and we will call it... `id` (mind blowing, right?):

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)

(dat %<>% mutate(id = school*10000 + photoid)) %>%
  head %>%
  select(school, photoid, id)
```

```
##   school photoid      id
```

```
## 1    111       1 1110001
## 2    111       2 1110002
## 3    111       7 1110007
## 4    111      13 1110013
## 5    111      14 1110014
## 6    111      15 1110015
```

Wow, what happend in the last three lines of code! What is that `%>%`? Well, that's the piping operator, and it is a very nice way of writing nested function calls. In this case, instead of having write something like

```
dat_filtered$id <- dat_filtered$school*10000 + dat_filtered$photoid
subset(head(dat_filtered), select = c(school, photoid, id))
```

4.2 Creating a network

- We want to build a social network. For that, we either use an adjacency matrix or an edgelist.
- Each individual of the SNS data nomitated 19 friends from school. We will use those nominations to create the social network.
- In this case, we will create the network by coercing the dataset into an edgelist.

4.2.1 From survey to edgelist

Let's start by loading a couple of handy R packages for this task, `tidyr` (Wickham and Henry 2017), which we will use to reshape the data, and `stringr` (Wickham 2017), which we will use to process strings using *regular expressions*[1].

```
library(tidyr)
library(stringr)
```

Optionally, we can use the `tibble` type of object which is an alternative to the actual `data.frame`. This object is claimed to provide *more efficient methods for matrices and data frames*.

```
dat <- as_tibble(dat)
```

What I like from tibbles is that when you print them on the console these actually look nice:

```
dat
```

```
## # A tibble: 2,164 x 100
##    photoid school hispanic female1 female2 female3 female4 grades1 grades2
```

[1] Please refer to the help file `?'regular expression'` in R. The R package rex (Ushey, Hester, and Krzyzanowski 2017) is a very nice companion for writing regular expressions. There's also a neat (but experimental) RStudio addin that can be very helpful for understanding how regular expressions work, the regexplain addin.

```
##      <int>  <int>    <dbl>   <int>   <int>   <int>   <int>   <dbl>   <dbl>
##  1       1    111        1      NA      NA       0       0     NA      NA
##  2       2    111        1       0      NA      NA       0      3       NA
##  3       7    111        0       1       1       1       1      5        4.5
##  4      13    111        1       1       1       1       1      2.5      2.5
##  5      14    111        1       1       1       1      NA      3        3.5
##  6      15    111        1       0       0       0       0      2.5      2.5
##  7      20    111        1       1       1       1       1      2.5      2.5
##  8      22    111        1      NA      NA       0       0     NA      NA
##  9      25    111        0       1       1      NA       1      4.5      3.5
## 10      27    111        1       0      NA       0       0      3.5     NA
## # ... with 2,154 more rows, and 91 more variables: grades3 <dbl>,
## #   grades4 <dbl>, eversmk1 <int>, eversmk2 <int>, eversmk3 <int>,
## #   eversmk4 <int>, everdrk1 <int>, everdrk2 <int>, everdrk3 <int>,
## #   everdrk4 <int>, home1 <int>, home2 <int>, home3 <int>, home4 <int>,
## #   sch_friend11 <int>, sch_friend12 <int>, sch_friend13 <int>,
## #   sch_friend14 <int>, sch_friend15 <int>, sch_friend16 <int>,
## #   sch_friend17 <int>, sch_friend18 <int>, sch_friend19 <int>,
## #   sch_friend110 <int>, sch_friend111 <int>, sch_friend112 <int>,
## #   sch_friend113 <int>, sch_friend114 <int>, sch_friend115 <int>,
## #   sch_friend116 <int>, sch_friend117 <int>, sch_friend118 <int>,
## #   sch_friend119 <int>, sch_friend21 <int>, sch_friend22 <int>,
## #   sch_friend23 <int>, sch_friend24 <int>, sch_friend25 <int>,
## #   sch_friend26 <int>, sch_friend27 <int>, sch_friend28 <int>,
## #   sch_friend29 <int>, sch_friend210 <int>, sch_friend211 <int>,
## #   sch_friend212 <int>, sch_friend213 <int>, sch_friend214 <int>,
## #   sch_friend215 <int>, sch_friend216 <int>, sch_friend217 <int>,
## #   sch_friend218 <int>, sch_friend219 <int>, sch_friend31 <int>,
## #   sch_friend32 <int>, sch_friend33 <int>, sch_friend34 <int>,
## #   sch_friend35 <int>, sch_friend36 <int>, sch_friend37 <int>,
## #   sch_friend38 <int>, sch_friend39 <int>, sch_friend310 <int>,
## #   sch_friend311 <int>, sch_friend312 <int>, sch_friend313 <int>,
## #   sch_friend314 <int>, sch_friend315 <int>, sch_friend316 <int>,
## #   sch_friend317 <int>, sch_friend318 <int>, sch_friend319 <int>,
## #   sch_friend41 <int>, sch_friend42 <int>, sch_friend43 <int>,
## #   sch_friend44 <int>, sch_friend45 <int>, sch_friend46 <int>,
## #   sch_friend47 <int>, sch_friend48 <int>, sch_friend49 <int>,
## #   sch_friend410 <int>, sch_friend411 <int>, sch_friend412 <int>,
## #   sch_friend413 <int>, sch_friend414 <int>, sch_friend415 <int>,
## #   sch_friend416 <int>, sch_friend417 <int>, sch_friend418 <int>,
## #   sch_friend419 <int>, id <dbl>
```

```r
# Maybe too much piping... but its cool!
net <- dat %>%
```

```
  select(id, school, starts_with("sch_friend")) %>%
  gather(key = "varname", value = "content", -id, -school) %>%
  filter(!is.na(content)) %>%
  mutate(
    friendid = school*10000 + content,
    year     = as.integer(str_extract(varname, "(?<=[a-z])[0-9]")),
    nnom     = as.integer(str_extract(varname, "(?<=[a-z][0-9])[0-9]+"))
  )
```

Let's take a look at this step by step:

1. First, we subset the data: We want to keep `id, school, sch_friend*`. For the later we use the function `starts_with` (from the `tidyselect` package). This allows us to select all variables that starts with the word "`sch_friend`," which means that `sch_friend11, sch_friend12, ...` will all be selected.

```
dat %>%
  select(id, school, starts_with("sch_friend"))
```

```
## # A tibble: 2,164 x 78
##         id school sch_friend11 sch_friend12 sch_friend13 sch_friend14
##      <dbl>  <int>        <int>        <int>        <int>        <int>
##  1 1.11e6    111           NA           NA           NA           NA
##  2 1.11e6    111          424          423          426          289
##  3 1.11e6    111          629          505           NA           NA
##  4 1.11e6    111          232          569           NA           NA
##  5 1.11e6    111          582          134           41          592
##  6 1.11e6    111           26          488           81          138
##  7 1.11e6    111          528           NA          492          395
##  8 1.11e6    111           NA           NA           NA           NA
##  9 1.11e6    111          135          185          553           84
## 10 1.11e6    111          346          168          559            5
## # ... with 2,154 more rows, and 72 more variables: sch_friend15 <int>,
## #   sch_friend16 <int>, sch_friend17 <int>, sch_friend18 <int>,
## #   sch_friend19 <int>, sch_friend110 <int>, sch_friend111 <int>,
## #   sch_friend112 <int>, sch_friend113 <int>, sch_friend114 <int>,
## #   sch_friend115 <int>, sch_friend116 <int>, sch_friend117 <int>,
## #   sch_friend118 <int>, sch_friend119 <int>, sch_friend21 <int>,
## #   sch_friend22 <int>, sch_friend23 <int>, sch_friend24 <int>,
## #   sch_friend25 <int>, sch_friend26 <int>, sch_friend27 <int>,
## #   sch_friend28 <int>, sch_friend29 <int>, sch_friend210 <int>,
## #   sch_friend211 <int>, sch_friend212 <int>, sch_friend213 <int>,
## #   sch_friend214 <int>, sch_friend215 <int>, sch_friend216 <int>,
## #   sch_friend217 <int>, sch_friend218 <int>, sch_friend219 <int>,
## #   sch_friend31 <int>, sch_friend32 <int>, sch_friend33 <int>,
```

```
## #   sch_friend34 <int>, sch_friend35 <int>, sch_friend36 <int>,
## #   sch_friend37 <int>, sch_friend38 <int>, sch_friend39 <int>,
## #   sch_friend310 <int>, sch_friend311 <int>, sch_friend312 <int>,
## #   sch_friend313 <int>, sch_friend314 <int>, sch_friend315 <int>,
## #   sch_friend316 <int>, sch_friend317 <int>, sch_friend318 <int>,
## #   sch_friend319 <int>, sch_friend41 <int>, sch_friend42 <int>,
## #   sch_friend43 <int>, sch_friend44 <int>, sch_friend45 <int>,
## #   sch_friend46 <int>, sch_friend47 <int>, sch_friend48 <int>,
## #   sch_friend49 <int>, sch_friend410 <int>, sch_friend411 <int>,
## #   sch_friend412 <int>, sch_friend413 <int>, sch_friend414 <int>,
## #   sch_friend415 <int>, sch_friend416 <int>, sch_friend417 <int>,
## #   sch_friend418 <int>, sch_friend419 <int>
```

2. Then, we reshape it to *long* format: By transposing all the `sch_friend*` to long. We do this by means of the function `gather` (from the `tidyr` package). This is an alternative to the `reshape` function, and I personally find it easier to use. Let's see how it works:

```
dat %>%
  select(id, school, starts_with("sch_friend")) %>%
  gather(key = "varname", value = "content", -id, -school)
```

```
## # A tibble: 164,464 x 4
##         id school varname      content
##      <dbl>  <int> <chr>          <int>
##  1 1110001    111 sch_friend11      NA
##  2 1110002    111 sch_friend11     424
##  3 1110007    111 sch_friend11     629
##  4 1110013    111 sch_friend11     232
##  5 1110014    111 sch_friend11     582
##  6 1110015    111 sch_friend11      26
##  7 1110020    111 sch_friend11     528
##  8 1110022    111 sch_friend11      NA
##  9 1110025    111 sch_friend11     135
## 10 1110027    111 sch_friend11     346
## # ... with 164,454 more rows
```

In this case the `key` parameter sets the name of the variable that will contain the name of the variable that was reshaped, while `value` is the name of the variable that will hold the content of the data (that's why I named those like that). The `-id, -school` bit tells the function to "drop" those variables before reshaping, in other words, "reshape everything but `id` and `school`."

Also, notice that we passed from 2164 rows to 19 (nominations) * 2164 (subjects) * 4 (waves) = 164464 rows, as expected.

3. As the nomination data can be empty for some cells, we need to take care of those cases, the `NA`s, so we filter the data:

```
dat %>%
  select(id, school, starts_with("sch_friend")) %>%
  gather(key = "varname", value = "content", -id, -school) %>%
  filter(!is.na(content))
```

```
## # A tibble: 39,561 x 4
##         id school varname      content
##      <dbl>  <int> <chr>          <int>
##  1 1110002    111 sch_friend11     424
##  2 1110007    111 sch_friend11     629
##  3 1110013    111 sch_friend11     232
##  4 1110014    111 sch_friend11     582
##  5 1110015    111 sch_friend11      26
##  6 1110020    111 sch_friend11     528
##  7 1110025    111 sch_friend11     135
##  8 1110027    111 sch_friend11     346
##  9 1110029    111 sch_friend11     369
## 10 1110030    111 sch_friend11     462
## # ... with 39,551 more rows
```

4. And finally, we create three new variables from this dataset: `friendid`, `year`, and `nom_num` (nomination number). All this using regular expressions:

```
dat %>%
  select(id, school, starts_with("sch_friend")) %>%
  gather(key = "varname", value = "content", -id, -school) %>%
  filter(!is.na(content)) %>%
  mutate(
    friendid = school*10000 + content,
    year     = as.integer(str_extract(varname, "(?<=[a-z])[0-9]")),
    nnom     = as.integer(str_extract(varname, "(?<=[a-z][0-9])[0-9]+"))
    )
```

```
## # A tibble: 39,561 x 7
##         id school varname      content friendid  year  nnom
##      <dbl>  <int> <chr>          <int>    <dbl> <int> <int>
##  1 1110002    111 sch_friend11     424  1110424     1     1
##  2 1110007    111 sch_friend11     629  1110629     1     1
##  3 1110013    111 sch_friend11     232  1110232     1     1
##  4 1110014    111 sch_friend11     582  1110582     1     1
##  5 1110015    111 sch_friend11      26  1110026     1     1
##  6 1110020    111 sch_friend11     528  1110528     1     1
##  7 1110025    111 sch_friend11     135  1110135     1     1
##  8 1110027    111 sch_friend11     346  1110346     1     1
##  9 1110029    111 sch_friend11     369  1110369     1     1
```

```
## 10 1110030    111 sch_friend11     462  1110462     1     1
## # ... with 39,551 more rows
```

The regular expression `(?<=[a-z])` matches a string that is preceeded by any letter from *a* to *z*, whereas the expression `[0-9]` matches a single number. Hence, from the string `"sch_friend12"`, the regular expression will only match the 1, as it is the only number followed by a letter. On the other hand, the expression `(?<=[a-z][0-9])` matches a string that is preceeded by a letter from *a* to *z* and a number from *0* to *9*; and the expression `[0-9]+` matches a string of numbers–so it could be more than one. Hence, from the string `"sch_friend12"`, we will get 2. We can actually se this

```
str_extract("sch_friend12", "(?<=[a-z])[0-9]")
```

```
## [1] "1"
```

```
str_extract("sch_friend12", "(?<=[a-z][0-9])[0-9]+")
```

```
## [1] "2"
```

And finally, the `as.integer` function coerces the returning value from the `str_extract` function from `character` to `integer`. Now that we have this edgelist, we can create an igraph object

4.2.2 igraph network

For coercing the edgelist into an igraph object, we will be using the `graph_from_data_frame` function in igraph (Csardi and Nepusz 2006). This function receives a data frame where the two first columns are sorce(ego) and target(alter), whether is it directed or not, and an optional data frame with vertices, in which's first column should contain the vertex ids.

Using the optional `vertices` argument is a good practice since by doing so you are telling the function what is the set of vertex ids that you are expecting to find. Using the original dataset, we will create a data frame name vertices:

```
vertex_attrs <- dat %>%
  select(id, school, hispanic, female1, starts_with("eversmk"))
```

Now, let's now use the function `graph_from_data_frame` to create an `igraph` object:

```
library(igraph)

ig_year1 <- net %>%
  filter(year == "1") %>%
  select(id, friendid, nnom) %>%
  graph_from_data_frame(
    vertices = vertex_attrs
  )
```

```
## Error in graph_from_data_frame(., vertices = vertex_attrs): Some vertex names in edge l
```

Ups! It seems that individuals are making nominations to other students that were not included on the survery. How to solve that? Well, it all depends on what you need to do! In this case, we will go for the *quietly-remove-em'-and-don't-tell* strategy:

```
ig_year1 <- net %>%
  filter(year == "1") %>%

  # Extra line, all nominations must be in ego too.
  filter(friendid %in% id) %>%

  select(id, friendid, nnom) %>%
  graph_from_data_frame(
    vertices = vertex_attrs
    )

ig_year1
```

```
## IGRAPH 71220e9 DN-- 2164 9514 --
## + attr: name (v/c), school (v/n), hispanic (v/n), female1 (v/n),
## | eversmk1 (v/n), eversmk2 (v/n), eversmk3 (v/n), eversmk4 (v/n),
## | nnom (e/n)
## + edges from 71220e9 (vertex names):
##  [1] 1110007->1110629 1110013->1110232 1110014->1110582 1110015->1110026
##  [5] 1110025->1110135 1110027->1110346 1110029->1110369 1110035->1110034
##  [9] 1110040->1110390 1110041->1110557 1110044->1110027 1110046->1110030
## [13] 1110050->1110086 1110057->1110263 1110069->1110544 1110071->1110167
## [17] 1110072->1110289 1110073->1110014 1110075->1110352 1110084->1110305
## [21] 1110086->1110206 1110093->1110040 1110094->1110483 1110095->1110043
## + ... omitted several edges
```

So there we have, our network with 2164 nodes and 9514 edges. The next steps: get some descriptive stats and visualize our network.

4.3 Network descriptive stats

While we could do all networks at once, in this part we will focus on computing some network statistics for one of the schools only. We start by school 111. The first question that you should be asking your self now is, "how can I get that information from the igraph object?." Well, vertex attributes and edges attributes can be accessed via the V and E functions respectively; moreover, we can list what vertex/edge attributes are available:

```
list.vertex.attributes(ig_year1)
```

```
## [1] "name"     "school"   "hispanic" "female1"  "eversmk1" "eversmk2" "eversmk3"
## [8] "eversmk4"
```

```
list.edge.attributes(ig_year1)
```

```
## [1] "nnom"
```

Just like we would do with data frames, accessing vertex attributes is done via the dollar sign operator `$` together with the `V` function, for example, accessing the first 10 elements of the variable `hispanic` can be done as follows:

```
V(ig_year1)$hispanic[1:10]
```

```
##  [1] 1 1 0 1 1 1 1 1 0 1
```

Now that you know how to access vertex attributes, we can get the network corresponding to school 111 by identifying which vertices are part of it and pass that information to the `induced_subgraph` function:

```
# Which ids are from school 111?
school111ids <- which(V(ig_year1)$school == 111)

# Creating a subgraph
ig_year1_111 <- induced_subgraph(
  graph = ig_year1,
  vids  = school111ids
)
```

The `which` function in R returns a vector of indices indicating which elements are true. In our case it will return a vector of indices of the vertices which have the attribute `school` equal to `111`. Now that we have our subgraph, we can compute different centrality measures[2] for each vertex and store them in the igraph object itself:

```
# Computing centrality measures for each vertex
V(ig_year1_111)$indegree   <- degree(ig_year1_111, mode = "in")
V(ig_year1_111)$outdegree  <- degree(ig_year1_111, mode = "out")
V(ig_year1_111)$closeness  <- closeness(ig_year1_111, mode = "total")
```

```
## Warning in closeness(ig_year1_111, mode = "total"): At centrality.c:
## 2784 :closeness centrality is not well-defined for disconnected graphs
```

```
V(ig_year1_111)$betweeness <- betweenness(ig_year1_111, normalized = TRUE)
```

From here, we can *go back* to our old habits and get the set of vertex attributes as a data frame so we can compute some summary statistics on the centrality measurements that we just got

```
# Extracting each vectex features as a data.frame
stats <- as_data_frame(ig_year1_111, what = "vertices")

# Computing quantiles for each variable
```

[2] For more information about the different centrality measurements, please take a look at the “Centrality” article on Wikipedia.

```
stats_degree <- with(stats, {
 cbind(
   indegree   = quantile(indegree, c(.025, .5, .975)),
   outdegree  = quantile(outdegree, c(.025, .5, .975)),
   closeness  = quantile(closeness, c(.025, .5, .975)),
   betweeness = quantile(betweeness, c(.025, .5, .975))
 )
})

stats_degree
```

```
##       indegree outdegree    closeness  betweeness
## 2.5%         0         0 3.526640e-06 0.000000000
## 50%          4         4 1.595431e-05 0.001879006
## 97.5%       16        16 1.601822e-05 0.016591048
```

The `with` function is somewhat similar to what `dplyr` allows us to do when we want to work with the dataset but without mentioning its name everytime that we ask for a variable. Without using the `with` function, the previous could have been done as follows:

```
stats_degree <-
 cbind(
   indegree   = quantile(stats$indegree, c(.025, .5, .975)),
   outdegree  = quantile(stats$outdegree, c(.025, .5, .975)),
   closeness  = quantile(stats$closeness, c(.025, .5, .975)),
   betweeness = quantile(stats$betweeness, c(.025, .5, .975))
 )
```

Now we will compute some statistics at the graph level:

```
cbind(
  size    = vcount(ig_year1_111),
  nedges  = ecount(ig_year1_111),
  density = edge_density(ig_year1_111),
  recip   = reciprocity(ig_year1_111),
  centr   = centr_betw(ig_year1_111)$centralization,
  pathLen = mean_distance(ig_year1_111)
  )
```

```
##      size nedges     density     recip      centr pathLen
## [1,]  533   2638 0.009303277 0.3731513 0.02179154 4.23678
```

Triadic census

```
triadic <- triad_census(ig_year1_111)
triadic
```

```
## [1] 24059676   724389   290849     3619     3383     4401     3219     2997
## [9]      407       33      836      235      163      137      277       85
```

To get a nicer view of this, we can use a table that I retrieved from `?triad_census`. Moreover, instead of looking a the raw counts, we can normalize the `triadic` object by its sum so we get proportions instead[3]

```
knitr::kable(cbind(
  Pcent = triadic/sum(triadic)*100,
  read.csv("triadic_census.csv")
  ), digits = 2)
```

Pcent	code	description
95.88	003	A,B,C, the empty graph.
2.89	012	A->B, C, the graph with a single directed edge.
1.16	102	A<->B, C, the graph with a mutual connection between two vertices.
0.01	021D	A<-B->C, the out-star.
0.01	021U	A->B<-C, the in-star.
0.02	021C	A->B->C, directed line.
0.01	111D	A<->B<-C.
0.01	111U	A<->B->C.
0.00	030T	A->B<-C, A->C.
0.00	030C	A<-B<-C, A->C.
0.00	201	A<->B<->C.
0.00	120D	A<-B->C, A<->C.
0.00	120U	A->B<-C, A<->C.
0.00	120C	A->B->C, A<->C.
0.00	210	A->B<->C, A<->C.
0.00	300	A<->B<->C, A<->C, the complete graph.

4.4 Plotting the network in igraph

4.4.1 Single plot

Let's take a look at how does our network looks like when we use the default parameters in the plot method of the igraph object:

```
plot(ig_year1)
```

Not very nice, right? A couple of things with this plot:

1. We are looking at all schools simultaneously, which does not make sense. So, instead of plotting `ig_year1`, we will focus on `ig_year1_111`.
2. All the vertices have the same size, and more over, are overalaping. So, instead of using

[3]During our workshop, Prof. De la Haye suggested using $\binom{n}{3}$ as a normalizing constant. It turns out that `sum(triadic) = choose(n, 3)`! So either approach is correct.

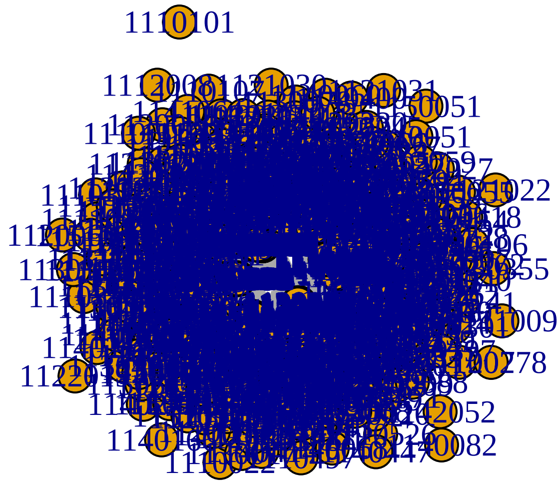


Figure 4.1: A not very nice network plot. This is what we get with the default parameters in igraph.

the default size, we will size the vertices by indegree using the `degree` function, and passing the vector of degrees to `vertex.size`.[4]

3. Given the number of vertices in these networks, the labels are not useful here. So we will remove them by setting `vertex.label = NA`. Moreover, we will reduce the size of the arrows' tip by setting `edge.arrow.size = 0.25`.

4. And finally, we will set the color of each vertex to be a function of whether the individual is hispanic or not. For this last bit we need to go a bit more of programming:

```
col_hispanic <- V(ig_year1_111)$hispanic + 1
col_hispanic <- coalesce(col_hispanic, 3)
col_hispanic <- c("steelblue", "tomato", "white")[col_hispanic]
```

Line by line, we did the following:

1. The first line added one to all no NA values, so that the 0s (non-hispanic) turned to 1s and the 1s (hispanic) turned to 2s.

2. The second line replaced all NAs with the number 3, so that our vector `col_hispanic` now ranges from 1 to 3 with no NAs in it.

3. In the last line we created a vector of colors. Essentially, what we are doing here is telling R to create a vector of length `length(col_hispanic)` by selecting elements by index from the vector `c("steelblue", "tomato", "white")`. This way, if, for example, the first element of the vector `col_hispanic` was a 3, our new vector of colors would have a `"white"` in it.

To make sure we know we are right, let's print the first 10 elements of our new vector of colors

[4] Figuring out what is the optimal vertex size is a bit tricky. Without getting too technical, there's no other way of getting *nice* vertex size other than just playing with different values of it. A nice solution to this is using `netdiffuseR::igraph_vertex_rescale` which rescales the vertices so that these keep their aspect ratio to a predefined proportion of the screen.

together with the original `hispanic` column:

```
cbind(
  original = V(ig_year1_111)$hispanic[1:10],
  colors   = col_hispanic[1:10]
  )
```

```
##       original colors
##  [1,] "1"      "tomato"
##  [2,] "1"      "tomato"
##  [3,] "0"      "steelblue"
##  [4,] "1"      "tomato"
##  [5,] "1"      "tomato"
##  [6,] "1"      "tomato"
##  [7,] "1"      "tomato"
##  [8,] "1"      "tomato"
##  [9,] "0"      "steelblue"
## [10,] "1"      "tomato"
```

With our nice vector of colors, now we can pass it to `plot.igraph` (which we call implicitly by just calling `plot`), via the `vertex.color` argument:

```
# Fancy graph
set.seed(1)
plot(
  ig_year1_111,
  vertex.size     = degree(ig_year1_111)/10 +1,
  vertex.label    = NA,
  edge.arrow.size = .25,
  vertex.color    = col_hispanic
  )
```

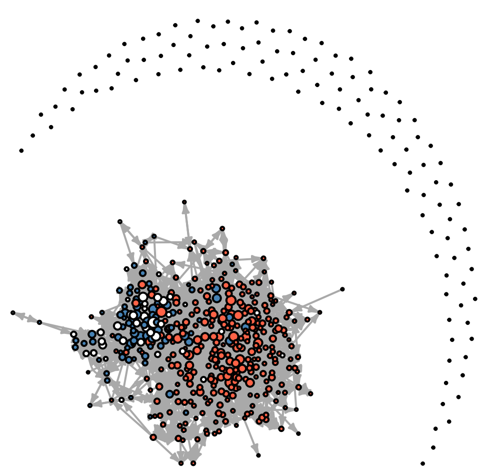


Figure 4.2: Friends network in time 1 for school 111.

Nice! So it does look better. The only problem is that we have a lot of isolates. Let's try again by drawing the same plot without isolates. To do so we need to filter the graph, for which we will use the function `induced_subgraph`

```
# Which vertices are not isolates?
which_ids <- which(degree(ig_year1_111, mode = "total") > 0)

# Getting the subgraph
ig_year1_111_sub <- induced_subgraph(ig_year1_111, which_ids)

# We need to get the same subset in col_hispanic
col_hispanic <- col_hispanic[which_ids]
```

```
# Fancy graph
set.seed(1)
plot(
  ig_year1_111_sub,
  vertex.size     = degree(ig_year1_111_sub)/5 +1,
  vertex.label    = NA,
  edge.arrow.size = .25,
  vertex.color    = col_hispanic
  )
```

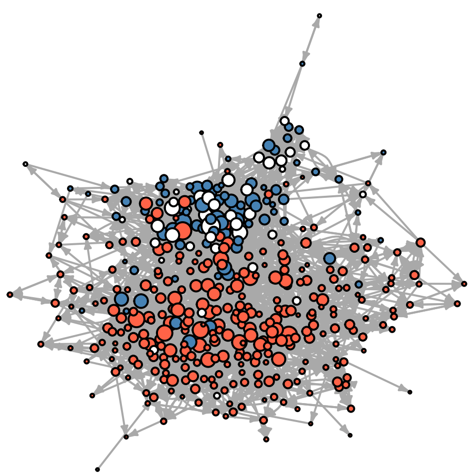


Figure 4.3: Friends network in time 1 for school 111. The graph excludes isolates.

Now that's better! An interesting pattern that shows up is that individuals seem to cluster by whether they are hispanic or not.

We can actually write this as a function so that, instead of us copying and pasting the code *n* times (supposing that we want to crate a plot similar to this *n* times). The next subsection does that.

4.4.2 Multiple plots

When you are repeating yourself over and over again, it is a good idea to write down a sequence of commands as a function. In this case, since we will be running the same type of plot for all schools/waves, we write a function in which the only things that changes are: (a) the school id, and (b) the color of the nodes.

```
myplot <- function(
  net,
  schoolid,
  mindgr = 1,
  vcol   = "tomato",
  ...) {

  # Creating a subgraph
  subnet <- induced_subgraph(
    net,
    which(degree(net, mode = "all") >= mindgr & V(net)$school == schoolid)
  )

  # Fancy graph
  set.seed(1)
  plot(
    subnet,
    vertex.size     = degree(subnet)/5,
    vertex.label    = NA,
    edge.arrow.size = .25,
    vertex.color    = vcol,
    ...
    )
}
```

The function definition:

1. The `myplot <- function([arguments]) {[body of the function]}` tells R that we are going to create a function called `myplot`.

2. In the arguments part, we are declaring 4 specific arguments: `net`, `schoolid`, `mindgr`, and `vcol`. These are an igraph object, the school id, the minimum degree that a vertex must have to be included in the plot, and the color of the vertices. Notice that, as a difference from other programming languages, in R we don't need to declare the types that these objects are.

3. The elipsis object, `...`, is a special object in R that allows us passing other arguments without us specifying which. In our case, if you take a look at the `plot` bit of the body of the function, you will see that we also added `...`; this means that whatever other

arguments (different from the ones that we explicitly defined) are passed to the function, these will be passed to the function `plot`, moreover, to the `plot.gexf` function (since the `subnet` object is actually an igraph object). In practice, this implies that we can, for example, set the argument `edge.arrow.size` when calling `myplot`, even though we did not included it in the function definition! (See `?dotsMethods` in R for more details).

In the following lines of code, using our new function, we will plot each schools' network in the same plotting device (window) with the help of the `par` function, and add legend with the `legend`:

```
# Plotting all together
oldpar <- par(no.readonly = TRUE)
par(mfrow = c(2, 3), mai = rep(0, 4), oma= c(1, 0, 0, 0))
myplot(ig_year1, 111, vcol = "tomato")
myplot(ig_year1, 112, vcol = "steelblue")
myplot(ig_year1, 113, vcol = "black")
myplot(ig_year1, 114, vcol = "gold")
myplot(ig_year1, 115, vcol = "white")
par(oldpar)

# A fancy legend
legend(
  "bottomright",
  legend = c(111, 112, 113, 114, 115),
  pt.bg  = c("tomato", "steelblue", "black", "gold", "white"),
  pch    = 21,
  cex    = 1,
  bty    = "n",
  title  = "School"
  )
```

So what happend here?

- `oldpar <- par(no.readonly = TRUE)` This line stores the current parameters for plotting. Since we are going to be changing them, we better make sure we are able to go back!.
- `par(mfrow = c(2, 3), mai = rep(0, 4), oma=rep(0, 4))` Here we are setting various things at the same time. `mfrow` specifies how many *figures* will be drawn and in what order, in particular, we are asking the plotting device to allow for 2*3 = 6 plots organized in 2 rows and 3 columns, and these will be drawn by row.

  `mai` specifies the size of the margins in inches. Setting all margins equal to zero (which is what we are doing now) gives more space to the network itself. The same is true for `oma`. See `?par` for more info.
- `myplot(ig_year1, ...)` This is simply calling our plotting function. The neat part of this

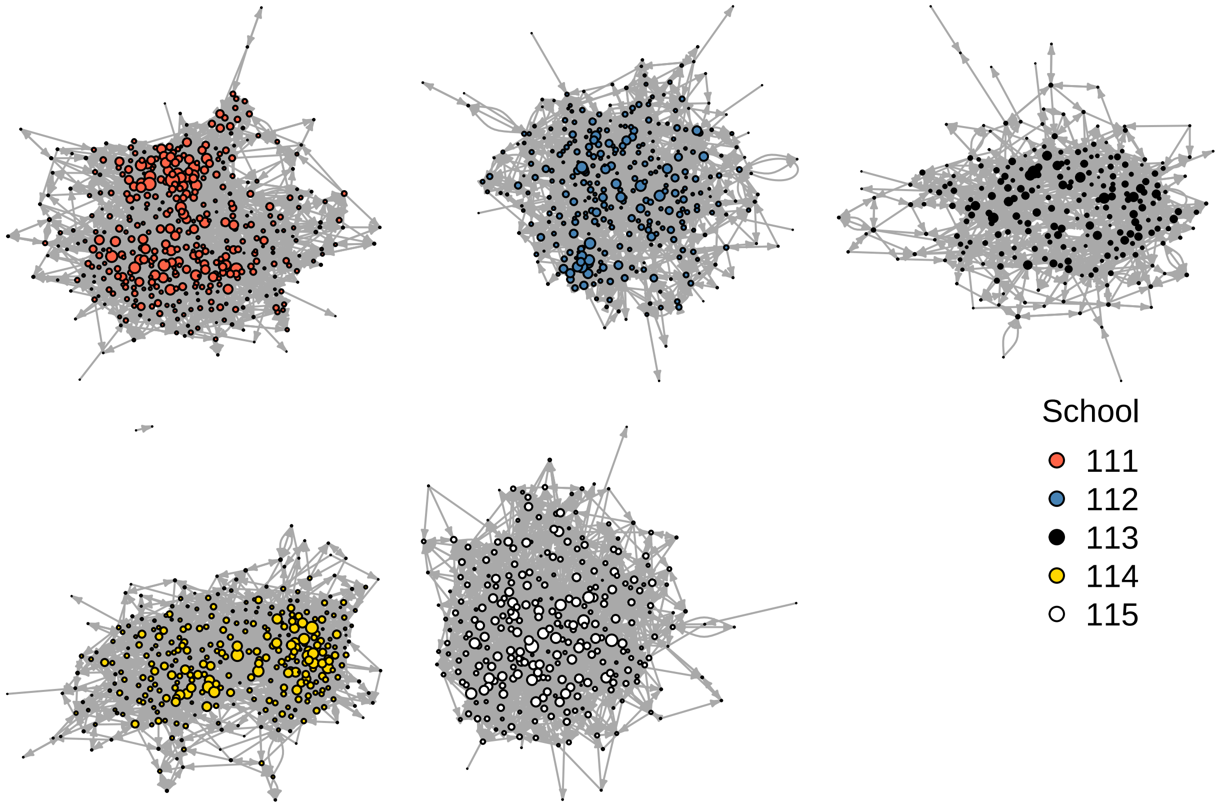


Figure 4.4: All 5 schools in time 1. Again, the graphs exclude isolates.

is that, since we set `mfrow = c(2, 3)`, R takes care of *distributing* the plots in the device.

- `par(oldpar)` This line allows us to restore the plotting parameters.

4.5 Statistical tests

4.5.1 Is nomination number correlated with indegree?

Hypothesis: Individuals that on average are among the first nominations of their peers are more popular

```
# Getting all the data in long format
edgelist <- as_long_data_frame(ig_year1) %>%
  as_tibble

# Computing indegree (again) and average nomination number
# Include "On a scale from one to five how close do you feel"
# Also for egocentric friends (A. Friends)
indeg_nom_cor <- group_by(edgelist, to, to_name, to_school) %>%
  summarise(
    indeg   = length(nnom),
```

```r
    nom_avg = 1/mean(nnom)
  ) %>%
  rename(
    school = to_school
  )

indeg_nom_cor
```

```
## # A tibble: 1,561 x 5
## # Groups:   to, to_name [1,561]
##        to to_name school indeg nom_avg
##     <dbl> <chr>    <int> <int>   <dbl>
##  1      2 1110002    111    22   0.222
##  2      3 1110007    111     7   0.175
##  3      4 1110013    111     6   0.171
##  4      5 1110014    111    19   0.134
##  5      6 1110015    111     3   0.15 
##  6      7 1110020    111     6   0.154
##  7      9 1110025    111     6   0.214
##  8     10 1110027    111    13   0.220
##  9     11 1110029    111    14   0.131
## 10     12 1110030    111     6   0.222
## # ... with 1,551 more rows
```

```r
# Using pearson's correlation
with(indeg_nom_cor, cor.test(indeg, nom_avg))
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  indeg and nom_avg
## t = -12.254, df = 1559, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3409964 -0.2504653
## sample estimates:
##        cor 
## -0.2963965
```

```r
save.image("03.rda")
```

Chapter 5

Exponential Random Graph Models

I strongly suggest reading the vignette included in the `ergm` R package

```
vignette("ergm", package="ergm")
```

So what are ERGMs anyway...

> The purpose of ERGMs, in a nutshell, is to describe parsimoniously the local selection forces that shape the global structure of a network. To this end, a network dataset, like those depicted in Figure 1, may be considered like the response in a regression model, where the predictors are things like "propensity for individuals of the same sex to form partnerships" or "propensity for individuals to form triangles of partnerships." In Figure 1(b), for example, it is evident that the individual nodes appear to cluster in groups of the same numerical labels (which turn out to be students' grades, 7 through 12); thus, an ERGM can help us quantify the strength of this intra-group effect.
>
> — (David R. Hunter et al. 2008)

The distribution of $\mathbf{Y}$ can be parameterized in the form

$$
\Pr\left(\mathbf{Y} = \mathbf{y}|\theta, \mathcal{Y}\right) = \frac{\exp\left\{\theta^{\mathsf{T}}\mathbf{g}(\mathbf{y})\right\}}{\kappa\left(\theta, \mathcal{Y}\right)},\quad \mathbf{y}\in\mathcal{Y}
$$

Where $\theta \in \Omega \subset \mathbb{R}^q$ is the vector of model coefficients and $\mathbf{g}(\mathbf{y})$ is a q-vector of statistics based on the adjacency matrix $\mathbf{y}$.

Model (5) may be expanded by replacing $\mathbf{g}(\mathbf{y})$ with $\mathbf{g}(\mathbf{y}, \mathbf{X})$ to allow for additional covariate information $\mathbf{X}$ about the network. The denominator,

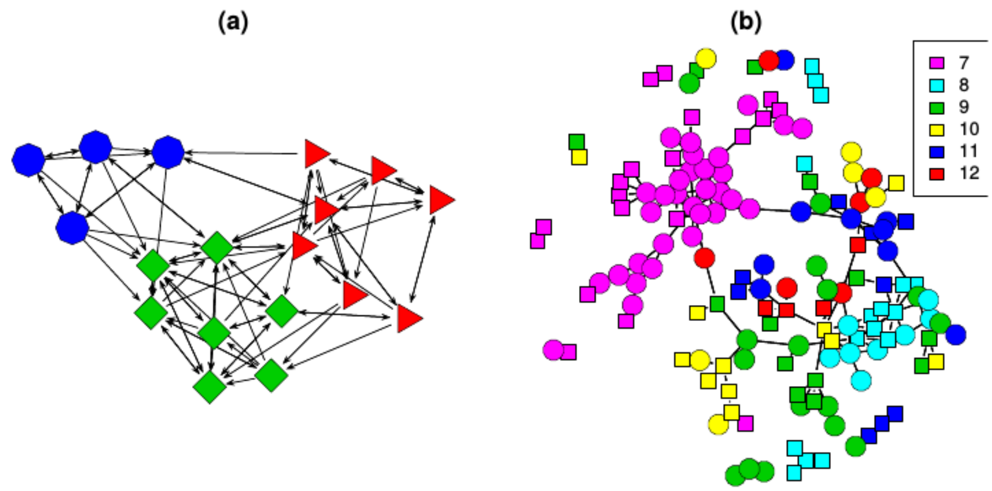


Figure 1: The **(a)** `samplike` and **(b)** `faux.mesa.high` networks described in Section 2. The values of nodal covariates may be indicated using various colors, shapes, and labels of nodes.

Figure 5.1: Source: Hunter et al. (2008)

$$\kappa(\theta, \mathcal{Y}) = \sum_{\mathbf{y}\in\mathcal{Y}} \exp\left\{\theta^{\mathsf{T}}\mathbf{g}(\mathbf{y})\right\}$$

Is the normalizing factor that ensures that equation (5) is a legitimate probability distribution. Even after fixing $\mathcal{Y}$ to be all the networks that have size n, the size of $\mathcal{Y}$ makes this type of models hard to estimate as there are $N = 2^{n(n-1)}$ possible networks! (David R. Hunter et al. 2008)

Recent developments include new forms of dependency structures, to take into account more general neighborhood effects. These models relax the one-step Markovian dependence assumptions, allowing investigation of longer range configurations, such as longer paths in the network or larger cycles (Pattison and Robins 2002). Models for bipartite (Faust and Skvoretz 1999) and tripartite (Mische and Robins 2000) network structures have also been developed. (David R. Hunter et al. 2008, 9)

5.0.1 A naive example

In the simplest case, ergm is equivalent to a logistic regression

```
library(ergm)
```

```
## Loading required package: network
```

```
## network: Classes for Relational Data
## Version 1.16.1 created on 2020-10-06.
## copyright (c) 2005, Carter T. Butts, University of California-Irvine
##                     Mark S. Handcock, University of California -- Los Angeles
```

```
##                      David R. Hunter, Penn State University
##                      Martina Morris, University of Washington
##                      Skye Bender-deMoll, University of Washington
##  For citation information, type citation("network").
##  Type help("network-package") to get started.

##
## ergm: version 3.11.0, created on 2020-10-14
## Copyright (c) 2020, Mark S. Handcock, University of California -- Los Angeles
##                      David R. Hunter, Penn State University
##                      Carter T. Butts, University of California -- Irvine
##                      Steven M. Goodreau, University of Washington
##                      Pavel N. Krivitsky, UNSW Sydney
##                      Martina Morris, University of Washington
##                      with contributions from
##                      Li Wang
##                      Kirk Li, University of Washington
##                      Skye Bender-deMoll, University of Washington
##                      Chad Klumb
##                      Michał Bojanowski, Kozminski University
##                      Ben Bolker
## Based on "statnet" project software (statnet.org).
## For license and citation information see statnet.org/attribution
## or type citation("ergm").

## NOTE: Versions before 3.6.1 had a bug in the implementation of the bd()
## constraint which distorted the sampled distribution somewhat. In
## addition, Sampson's Monks datasets had mislabeled vertices. See the
## NEWS and the documentation for more details.

## NOTE: Some common term arguments pertaining to vertex attribute and
## level selection have changed in 3.10.0. See terms help for more
## details. Use 'options(ergm.term=list(version="3.9.4"))' to use old
## behavior.
```

```
data("sampson")

samplike
```

```
##  Network attributes:
##   vertices = 18
##   directed = TRUE
##   hyper = FALSE
##   loops = FALSE
##   multiple = FALSE
##   total edges= 88
```

```
##     missing edges= 0
##     non-missing edges= 88
##
##  Vertex attribute names:
##     cloisterville group vertex.names
##
##  Edge attribute names:
##     nominations
```

```
y <- sort(as.vector(as.matrix(samplike)))[-c(1:18)]
glm(y~1, family=binomial("logit"))
```

```
##
## Call:  glm(formula = y ~ 1, family = binomial("logit"))
##
## Coefficients:
## (Intercept)
##     -0.9072
##
## Degrees of Freedom: 305 Total (i.e. Null);  305 Residual
## Null Deviance:       367.2
## Residual Deviance: 367.2     AIC: 369.2
```

```
ergm(samplike ~ edges)
```

```
## Starting maximum pseudolikelihood estimation (MPLE):

## Evaluating the predictor and response matrix.

## Maximizing the pseudolikelihood.

## Finished MPLE.

## Stopping at the initial estimate.

## Evaluating log-likelihood at the estimate.

##
## Call:
## ergm(formula = samplike ~ edges)
##
##
## MLE Coefficients:
##   edges
## -0.9072
```

```
pr <- mean(y)
log(pr) - log(1-pr) # Logit function
```

```
## [1] -0.9071582
```

```
qlogis(pr)
```

```
## [1] -0.9071582
```

5.1 Estimation of ERGMs

The ultimate goal is to be able to do statistical inference on the proposed model. In a *normal* setting, we would be able to use Maximum-Likelihood-Estimation (MLE) which basically consists on finding the model parameters θ that, given the observed data, maximizes the likelihood of the model. Such is usually done by applying Newton's method which requires been able to compute the log-likelihood of the model. This is a bit more complicated in ERGMs.

In the case of ERGMs, since part of the likelihood involves a normalizing constant that is a function of all possible networks, this is not as straight forward as it is in the regular setting. This is why we rely on simulations.

In `statnet`, the default estimation method is based on a method proposed by (Geyer and Thompson 1992), Markov-Chain MLE, which uses Markov-Chain Monte Carlo for simulating networks and a modified version of the Newton-Raphson algorihm to do the paremeter estimation part.

The idea of MC-MLE for this family of statistical models is the fact that the expectation of normalizing constant ratios can be approximated using the law of large numbers. In particular, the following:

$$
\begin{aligned}
\frac{\kappa(\theta,\mathcal{Y})}{\kappa(\theta_0,\mathcal{Y})} &= \frac{\sum_{\mathbf{y}\in\mathcal{Y}}\exp\left\{\theta^{\mathsf{T}}\mathbf{g}(\mathbf{y})\right\}}{\sum_{\mathbf{y}\in\mathcal{Y}}\exp\left\{\theta_0^{\mathsf{T}}\mathbf{g}(\mathbf{y})\right\}} \\
&= \sum_{\mathbf{y}\in\mathcal{Y}}\left(\frac{1}{\sum_{\mathbf{y}\in\mathcal{Y}}\exp\left\{\theta_0^{\mathsf{T}}\mathbf{g}(\mathbf{y})\right\}}\times\exp\left\{\theta^{\mathsf{T}}\mathbf{g}(\mathbf{y})\right\}\right) \\
&= \sum_{\mathbf{y}\in\mathcal{Y}}\left(\frac{\exp\left\{\theta_0^{\mathsf{T}}\mathbf{g}(\mathbf{y})\right\}}{\sum_{\mathbf{y}\in\mathcal{Y}}\exp\left\{\theta_0^{\mathsf{T}}\mathbf{g}(\mathbf{y})\right\}}\times\exp\left\{(\theta-\theta_0)^{\mathsf{T}}\mathbf{g}(\mathbf{y})\right\}\right) \\
&= \sum_{\mathbf{y}\in\mathcal{Y}}\left(\Pr(Y=y|\mathcal{Y},\theta_0)\times\exp\left\{(\theta-\theta_0)^{\mathsf{T}}\mathbf{g}(\mathbf{y})\right\}\right) \\
&= \mathrm{E}_{\theta_0}\left(\exp\left\{(\theta-\theta_0)^{\mathsf{T}}\mathbf{g}(\mathbf{y})\right\}\right)
\end{aligned}
$$

The final line can be approximated by the law of large numbers. In particular, the MC-MLE algorithm uses this fact to maximize the ratio of log-likelihoods. The objective function can be approximated by simulating m networks from the distribution with parameter θ_0:

$$
l(\theta)-l(\theta_0)\approx(\theta-\theta_0)^{\mathsf{T}}\mathbf{g}(\mathbf{y}_{obs})-\log\left[\frac{1}{m}\sum_{i=1}^{m}\exp\left\{(\theta-\theta_0)^{\mathsf{T}}\right\}\mathbf{g}(\mathbf{Y}_i)\right]
$$

For more details see (David R. Hunter et al. 2008). A sketch of the algorithm follows:

1. Initialize the algorithm with an initial guess of θ, call it $\theta^{(t)}$ (must be a rather OK guess)
2. While (no convergence) do:
   a. Using $\theta^{(t)}$, simulate M networks by means of small changes in the $\mathbf{Y}_{obs}$ (the observed network). This part is done by using an importance-sampling method which weights each proposed network by it's likelihood conditional on $\theta^{(t)}$
   b. With the networks simulated, we can do the Newton step to update the parameter $\theta^{(t)}$ (this is the iteration part in the `ergm` package): $\theta^{(t)} \to \theta^{(t+1)}$.
   c. If convergence has been reach (which usually means that $\theta^{(t)}$ and $\theta^{(t+1)}$ are not very different), then stop, otherwise, go to step a.

For more details see (Lusher, Koskinen, and Robins 2012; Admiraal and Handcock 2006; T. A. Snijders 2002; Wang et al. 2009) provides details on the algorithm used by PNet (which is the same as the one used in `RSiena`). (Lusher, Koskinen, and Robins 2012) provides a short discussion on differences between `ergm` and `PNet`.

5.2 The `ergm` package

The `ergm` R package (Handcock et al. 2017)

From the previous section:[1]

```
library(igraph)
library(magrittr)
library(dplyr)

load("03.rda")
```

In this section we will use the `ergm` package (from the `statnet` suit of packages (Handcock et al. 2016)) suit, and the `intergraph` (Bojanowski 2015) package. The latter provides functions to go back and forth between `igraph` and `network` objects from the `igraph` and `network` packages respectively[2]

```
library(ergm)
library(intergraph)
```

As a rather important side note, the order in which R packages are loaded matters. Why is this important to mention now? Well, it turns out that at least a couple of functions in the `network` package have the same name of some functions in the `igraph` package. When the `ergm` package is loaded, since it depends on `network`, it will load the `network` package first, which will *mask* some functions in `igraph`. This becomes evident once you load `ergm` after loading `igraph`:

[1] You can download the 03.rda file from this link.
[2] Yes, the classes have the same name as the packages.

```
The following objects are masked from 'package:igraph':

  add.edges, add.vertices, %c%, delete.edges, delete.vertices, get.edge.attribute, get.edg
  get.vertex.attribute, is.bipartite, is.directed, list.edge.attributes, list.vertex.attri
  set.edge.attribute, set.vertex.attribute
```

What are the implications of this? If you call the function `list.edge.attributes` for an object of class `igraph` R will return an error as the first function that matches that name comes from the `network` package! To avoid this you can use the double colon notation:

```
igraph::list.edge.attributes(my_igraph_object)
network::list.edge.attributes(my_network_object)
```

Anyway... Using the `asNetwork` function, we can coerce the igraph object into a network object so we can use it with the `ergm` function:

```
# Creating the new network
network_111 <- intergraph::asNetwork(ig_year1_111)

# Running a simple ergm (only fitting edge count)
ergm(network_111 ~ edges)
```

```
## [1] "Warning:  This network contains loops"
## [1] "Warning:  This network contains loops"

## Starting maximum pseudolikelihood estimation (MPLE):

## Evaluating the predictor and response matrix.

## Maximizing the pseudolikelihood.

## Finished MPLE.

## Stopping at the initial estimate.

## Evaluating log-likelihood at the estimate.

##
## MLE Coefficients:
##  edges
## -4.734
```

So what happened here! We got a warning. It turns out that our network has loops (didn't thought about it before!). Let's take a look on that with the `which_loop` function

```
E(ig_year1_111)[which_loop(ig_year1_111)]
```

```
## + 1/2638 edge from afe0cc1 (vertex names):
## [1] 1110111->1110111
```

We can get rid of these using the `igraph::-.igraph`. Moreover, just to illustrate how it can be done, let's get rid of the isolates using the same operator

```r
# Creating the new network
network_111 <- ig_year1_111

# Removing loops
network_111 <- network_111 - E(network_111)[which(which_loop(network_111))]

# Removing isolates
network_111 <- network_111 - which(degree(network_111, mode = "all") == 0)

# Converting the network
network_111 <- intergraph::asNetwork(network_111)
```

`asNetwork(simplify(ig_year1_111))` `ig_year1_111 %>% simplify %>% asNetwork`

A problem that we have on this data is the fact that some vertices have missing values in the variables `hispanic`, `female1`, and `eversmk1`. For now, we will proceed by imputing values based on the avareges:

```r
for (v in c("hispanic", "female1", "eversmk1")) {
  tmpv <- network_111 %v% v
  tmpv[is.na(tmpv)] <- mean(tmpv, na.rm = TRUE) > .5
  network_111 %v% v <- tmpv
}
```

5.3 Running ERGMs

Proposed workflow:

1. Estimate the simplest model, adding one variable at a time.
2. After each estimation, run the `mcmc.diagnostics` function to see how good/bad behaved are the chains.
3. Run the `gof` function to see how good is the model at matching the network's structural statistics.

What to use:

1. `control.ergms`: Maximum number of iteration, seed for Pseudo-RNG, how many cores
2. `ergm.constraints`: Where to sample the network from. Gives stability and (in some cases) faster convergence as by constraining the model you are reducing the sample size.

Here is an example of a couple of models that we could compare[3]

[3]Notice that this document may not include the usual messages that the `ergm` command generates during the estimation procedure. This is just to make it more printable-friendly.

```r
ans0 <- ergm(
  network_111 ~
    edges +
    nodematch("hispanic") +
    nodematch("female1") +
    nodematch("eversmk1") +
    mutual
    ,
  constraints = ~bd(maxout = 19),
  control = control.ergm(
    seed        = 1,
    MCMLE.maxit = 10,
    parallel    = 4,
    CD.maxit    = 10
    )
  )
```

```
## Warning in nobs.ergm(object, ...): The number of observed dyads in this
## network is ill-defined due to complex constraints on the sample space.
## Disable this warning with 'options(ergm.loglik.warn_dyads=FALSE)'.

## Warning in nobs.ergm(object, ...): The number of observed dyads in this
## network is ill-defined due to complex constraints on the sample space.
## Disable this warning with 'options(ergm.loglik.warn_dyads=FALSE)'.
```

So what are we doing here:

1. The model is controling for:

    a. edges Number of edges in the network (as opposed to its density)

    b. `nodematch("some-variable-name-here")` Includes a term that controls for homophily/heterophily

    c. `mutual` Number of mutual connections between (i, j), (j, i). This can be related to, for example, triadic closure.

For more on control parameters, see (Morris, Handcock, and Hunter 2008).

```r
ans1 <- ergm(
  network_111 ~
    edges +
    nodematch("hispanic") +
    nodematch("female1") +
    nodematch("eversmk1")
    ,
  constraints = ~bd(maxout = 19),
```

```
  control = control.ergm(
    seed        = 1,
    MCMLE.maxit = 10,
    parallel    = 4,
    CD.maxit    = 10
    )
  )
```

```
## Warning in nobs.ergm(object, ...): The number of observed dyads in this
## network is ill-defined due to complex constraints on the sample space.
## Disable this warning with 'options(ergm.loglik.warn_dyads=FALSE)'.

## Warning in nobs.ergm(object, ...): The number of observed dyads in this
## network is ill-defined due to complex constraints on the sample space.
## Disable this warning with 'options(ergm.loglik.warn_dyads=FALSE)'.
```

This example takes longer to compute

```
ans2 <- ergm(
  network_111 ~
    edges +
    nodematch("hispanic") +
    nodematch("female1") +
    nodematch("eversmk1") +
    mutual +
    balance
    ,
  constraints = ~bd(maxout = 19),
  control = control.ergm(
    seed        = 1,
    MCMLE.maxit = 10,
    parallel    = 4,
    CD.maxit    = 10
    )
  )
```

```
## Warning in nobs.ergm(object, ...): The number of observed dyads in this
## network is ill-defined due to complex constraints on the sample space.
## Disable this warning with 'options(ergm.loglik.warn_dyads=FALSE)'.

## Warning in nobs.ergm(object, ...): The number of observed dyads in this
## network is ill-defined due to complex constraints on the sample space.
## Disable this warning with 'options(ergm.loglik.warn_dyads=FALSE)'.
```

Now, a nice trick to see all regressions in the same table, we can use the `texreg` package

(Leifeld 2013) which supports ergm ouputs!

```
library(texreg)
```

```
## Version:  1.37.5
## Date:     2020-06-17
## Author:   Philip Leifeld (University of Essex)
##
## Consider submitting praise using the praise or praise_interactive functions.
## Please cite the JSS article in your publications -- see citation("texreg").
```

```
##
## Attaching package: 'texreg'
```

```
## The following object is masked from 'package:magrittr':
##
##     extract
```

```
screenreg(list(ans0, ans1, ans2))
```

```
## Warning: This object was fit with 'ergm' version 3.10.4.5075 or earlier.
## Summarizing it with version 3.11 or later may return incorrect results or fail.

## Warning: This object was fit with 'ergm' version 3.10.4.5075 or earlier.
## Summarizing it with version 3.11 or later may return incorrect results or fail.

## Warning: This object was fit with 'ergm' version 3.10.4.5075 or earlier.
## Summarizing it with version 3.11 or later may return incorrect results or fail.
```

```
##
## ===============================================================
##                      Model 1        Model 2        Model 3
## ---------------------------------------------------------------
## edges                    -5.64 ***      -5.52 ***      -5.58 ***
##                          (0.05)         (0.06)         (0.06)
## nodematch.hispanic        0.36 ***       0.50 ***       0.40 ***
##                          (0.04)         (0.04)         (0.04)
## nodematch.female1         0.83 ***       1.10 ***       0.83 ***
##                          (0.04)         (0.05)         (0.04)
## nodematch.eversmk1        0.35 ***       0.46 ***       0.36 ***
##                          (0.04)         (0.05)         (0.04)
## mutual                    4.09 ***                     -3.55 ***
##                          (0.07)                        (0.25)
## balance                                                 0.02 ***
##                                                        (0.00)
## ---------------------------------------------------------------
## AIC                  -32986.67      -31399.10      -33035.32
```

	Model 1	Model 2	Model 3
edges	-5.64^{***}	-5.52^{***}	-5.58^{***}
	(0.05)	(0.06)	(0.06)
nodematch.hispanic	0.36^{***}	0.50^{***}	0.40^{***}
	(0.04)	(0.04)	(0.04)
nodematch.female1	0.83^{***}	1.10^{***}	0.83^{***}
	(0.04)	(0.05)	(0.04)
nodematch.eversmk1	0.35^{***}	0.46^{***}	0.36^{***}
	(0.04)	(0.05)	(0.04)
mutual	4.09^{***}		-3.55^{***}
	(0.07)		(0.25)
balance			0.02^{***}
			(0.00)
AIC	−32986.67	−31399.10	−33035.32
BIC	−32936.32	−31358.82	−32974.91
Log Likelihood	16498.33	15703.55	16523.66

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 5.1: Statistical models

```
## BIC                   -32936.32      -31358.82      -32974.91
## Log Likelihood         16498.33       15703.55       16523.66
## ===============================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

Or, if you are using rmarkdown, you can export the results using LaTeX or html, let's try the latter to see how it looks like here:

```
library(texreg)
texreg(list(ans0, ans1, ans2))
```

```
## Warning: This object was fit with 'ergm' version 3.10.4.5075 or earlier.
## Summarizing it with version 3.11 or later may return incorrect results or fail.

## Warning: This object was fit with 'ergm' version 3.10.4.5075 or earlier.
## Summarizing it with version 3.11 or later may return incorrect results or fail.

## Warning: This object was fit with 'ergm' version 3.10.4.5075 or earlier.
## Summarizing it with version 3.11 or later may return incorrect results or fail.
```

5.4 Model Goodness-of-Fit

In raw terms, once each chain has reach stationary distribution, we can say that there are no problems with autocorrelation and that each sample point is iid. This implies that, since we are running the model with more than 1 chain, we can use all the samples (chains) as a single dataset.

> Recent changes in the ergm estimation algorithm mean that these plots can no longer be used to ensure that the mean statistics from the model match the observed

> network statistics. For that functionality, please use the GOF command: gof(object, GOF=~model).
>
> —?ergm::mcmc.diagnostics

Since `ans0` is the one model which did best, let's take a look at it's GOF statistics. First, lets see how the MCMC did. For this we can use the `mcmc.diagnostics` function including in the package. This function is actually a wrapper of a couple of functions from the coda package (Plummer et al. 2006) which is called upon the `$sample` object which holds the *centered* statistics from the sampled networks. This last point is important to consider since at first look it can be confusing to look at the `$sample` object since it neither matches the observed statistics, nor the coefficients.

When calling the function `mcmc.diagnostics(ans0, centered = FALSE)`, you will see a lot of output including a couple of plots showing the trace and posterior distribution of the *uncentered* statistics (`centered = FALSE`). In the next code chunks we will reproduce the output from the `mcmc.diagnostics` function step by step using the coda package. First we need to *uncenter* the sample object:

```r
# Getting the centered sample
sample_centered <- ans0$sample

# Getting the observed statistics and turning it into a matrix so we can add it
# to the samples
observed <- summary(ans0$formula)
observed <- matrix(
  observed,
  nrow  = nrow(sample_centered[[1]]),
  ncol  = length(observed),
  byrow = TRUE
  )

# Now we uncenter the sample
sample_uncentered <- lapply(sample_centered, function(x) {
  x + observed
})

# We have to make it an mcmc.list object
sample_uncentered <- coda::mcmc.list(sample_uncentered)
```

This is what is called under the hood:

1. *Empirical means and sd, and quantiles*:

   ```r
   summary(sample_uncentered)
   ```

   ```
   ##
   ## Iterations = 16384:1063936
   ```

```
## Thinning interval = 1024
## Number of chains = 4
## Sample size per chain = 1024
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                        Mean    SD Naive SE Time-series SE
## edges              2474.3 55.40   0.8656          4.129
## nodematch.hispanic 1836.5 43.56   0.6806          3.836
## nodematch.female1  1867.3 49.50   0.7735          4.900
## nodematch.eversmk1 1755.0 45.32   0.7081          2.926
## mutual              485.1 20.07   0.3136          3.544
##
## 2. Quantiles for each variable:
##
##                    2.5%  25%  50%  75% 97.5%
## edges              2365 2438 2475 2511  2580
## nodematch.hispanic 1747 1807 1838 1867  1918
## nodematch.female1  1778 1833 1866 1898  1975
## nodematch.eversmk1 1664 1726 1755 1784  1841
## mutual              446  472  485  498   527
```

2. *Cross correlation*:

```
coda::crosscorr(sample_uncentered)
```

```
##                        edges nodematch.hispanic nodematch.female1
## edges              1.0000000          0.8099803         0.8419023
## nodematch.hispanic 0.8099803          1.0000000         0.6845240
## nodematch.female1  0.8419023          0.6845240         1.0000000
## nodematch.eversmk1 0.8127786          0.6668579         0.6946880
## mutual             0.7144121          0.6064003         0.6720229
##                    nodematch.eversmk1    mutual
## edges                       0.8127786 0.7144121
## nodematch.hispanic          0.6668579 0.6064003
## nodematch.female1           0.6946880 0.6720229
## nodematch.eversmk1          1.0000000 0.5909593
## mutual                      0.5909593 1.0000000
```

3. *Autocorrelation*: Just for now, we will only take a look at autocorrelation for chain 1 only. Autocorrelation should be rather low (in a general MCMC setting). If autocorrelation is high, then it means that your sample is not idd (no markov property). A way out to solve this is *thinning* the sample.

```r
coda::autocorr(sample_uncentered)[[1]]
```

```
## , , edges
##
##                edges nodematch.hispanic nodematch.female1 nodematch.eversmk1
## Lag 0     1.0000000          0.8139761         0.7795009          0.7837272
## Lag 1024  0.8868373          0.7222805         0.6971419          0.7006181
## Lag 5120  0.5948994          0.5251881         0.4922158          0.4847903
## Lag 10240 0.4600845          0.4504976         0.3755953          0.3949433
## Lag 51200 0.1982049          0.2079237         0.3221285          0.3131978
##               mutual
## Lag 0     0.6565207
## Lag 1024  0.6511008
## Lag 5120  0.6123226
## Lag 10240 0.5450916
## Lag 51200 0.3781621
##
## , , nodematch.hispanic
##
##                edges nodematch.hispanic nodematch.female1 nodematch.eversmk1
## Lag 0     0.8139761          1.0000000         0.6678498          0.5997413
## Lag 1024  0.7368708          0.8947521         0.6118379          0.5398809
## Lag 5120  0.5294057          0.6364242         0.4658087          0.3828159
## Lag 10240 0.4054664          0.4877295         0.3715878          0.2940047
## Lag 51200 0.2058656          0.1750285         0.3230496          0.2682960
##               mutual
## Lag 0     0.6338096
## Lag 1024  0.6235126
## Lag 5120  0.5759901
## Lag 10240 0.5148339
## Lag 51200 0.3923427
##
## , , nodematch.female1
##
##                edges nodematch.hispanic nodematch.female1 nodematch.eversmk1
## Lag 0     0.7795009          0.6678498         1.0000000          0.5886437
## Lag 1024  0.6998063          0.6046370         0.9102620          0.5273102
## Lag 5120  0.4930271          0.4699355         0.6838324          0.3701848
## Lag 10240 0.3680917          0.3863329         0.5241266          0.2933634
## Lag 51200 0.1291978          0.1212720         0.3078540          0.1884934
##               mutual
## Lag 0     0.6480628
## Lag 1024  0.6419102
```

```
## Lag 5120  0.6093541
## Lag 10240 0.5327467
## Lag 51200 0.3444436
##
## , , nodematch.eversmk1
##
##                edges nodematch.hispanic nodematch.female1 nodematch.eversmk1
## Lag 0     0.7837272          0.5997413         0.5886437          1.0000000
## Lag 1024  0.6948882          0.5391618         0.5277555          0.9024858
## Lag 5120  0.4488066          0.4103141         0.3543596          0.6426104
## Lag 10240 0.3440736          0.3622540         0.2786189          0.5235972
## Lag 51200 0.1413846          0.1251185         0.3037022          0.3427353
##              mutual
## Lag 0     0.5189905
## Lag 1024  0.5109281
## Lag 5120  0.4754632
## Lag 10240 0.4043018
## Lag 51200 0.2511635
##
## , , mutual
##
##                edges nodematch.hispanic nodematch.female1 nodematch.eversmk1
## Lag 0     0.6565207          0.6338096         0.6480628          0.5189905
## Lag 1024  0.6473638          0.6296240         0.6400673          0.5133709
## Lag 5120  0.6106484          0.6120531         0.6093092          0.4949412
## Lag 10240 0.5779115          0.6078153         0.5675734          0.4953194
## Lag 51200 0.3343059          0.3086253         0.4037995          0.4237535
##              mutual
## Lag 0     1.0000000
## Lag 1024  0.9825012
## Lag 5120  0.9123847
## Lag 10240 0.8212019
## Lag 51200 0.4968927
```

4. *Geweke Diagnostic*: From the function's help file:

   > "If the samples are drawn from the stationary distribution of the chain, the two means are equal and Geweke's statistic has an asymptotically standard normal distribution. [. . .] The Z-score is calculated under the assumption that the two parts of the chain are asymptotically independent, which requires that the sum of frac1 and frac2 be strictly less than 1.""
   >
   > —?coda::geweke.diag

   Let's take a look at a single chain:

```
coda::geweke.diag(sample_uncentered)[[1]]
```

```
##
## Fraction in 1st window = 0.1
## Fraction in 2nd window = 0.5
##
##              edges nodematch.hispanic  nodematch.female1 nodematch.eversmk1
##             0.5295             0.4904             1.6354             0.6644
##             mutual
##             1.1170
```

5. *(not included) Gelman Diagnostic*: From the function's help file:

    > Gelman and Rubin (1992) propose a general approach to monitoring convergence of MCMC output in which m > 1 parallel chains are run with starting values that are overdispersed relative to the posterior distribution. Convergence is diagnosed when the chains have 'forgotten' their initial values, and the output from all chains is indistinguishable. The gelman.diag diagnostic is applied to a single variable from the chain. It is based a comparison of within-chain and between-chain variances, and is similar to a classical analysis of variance. —?coda::gelman.diag

    As a difference from the previous diagnostic statistic, this uses all chains simulatenously:

    ```
    coda::gelman.diag(sample_uncentered)
    ```

    ```
    ## Potential scale reduction factors:
    ##
    ##                      Point est. Upper C.I.
    ## edges                      1.16       1.42
    ## nodematch.hispanic         1.10       1.28
    ## nodematch.female1          1.28       1.68
    ## nodematch.eversmk1         1.34       1.81
    ## mutual                     1.32       1.79
    ##
    ## Multivariate psrf
    ##
    ## 1.44
    ```

    As a rule of thumb, values that are in the $[.9, 1.1]$ are good.

One nice feature of the `mcmc.diagnostics` function is the nice trace and posterior distribution plots that it generates. If you have the R package `latticeExtra` (Sarkar and Andrews 2016), the function will override the default plots used by `coda::plot.mcmc` and use lattice instead, creating a nicer looking plots. The next code chunk calls the `mcmc.diagnostic` function, but we suppress the rest of the output (see figure **??**).

```
mcmc.diagnostics(ans0, center = FALSE) # Suppressing all the output
```

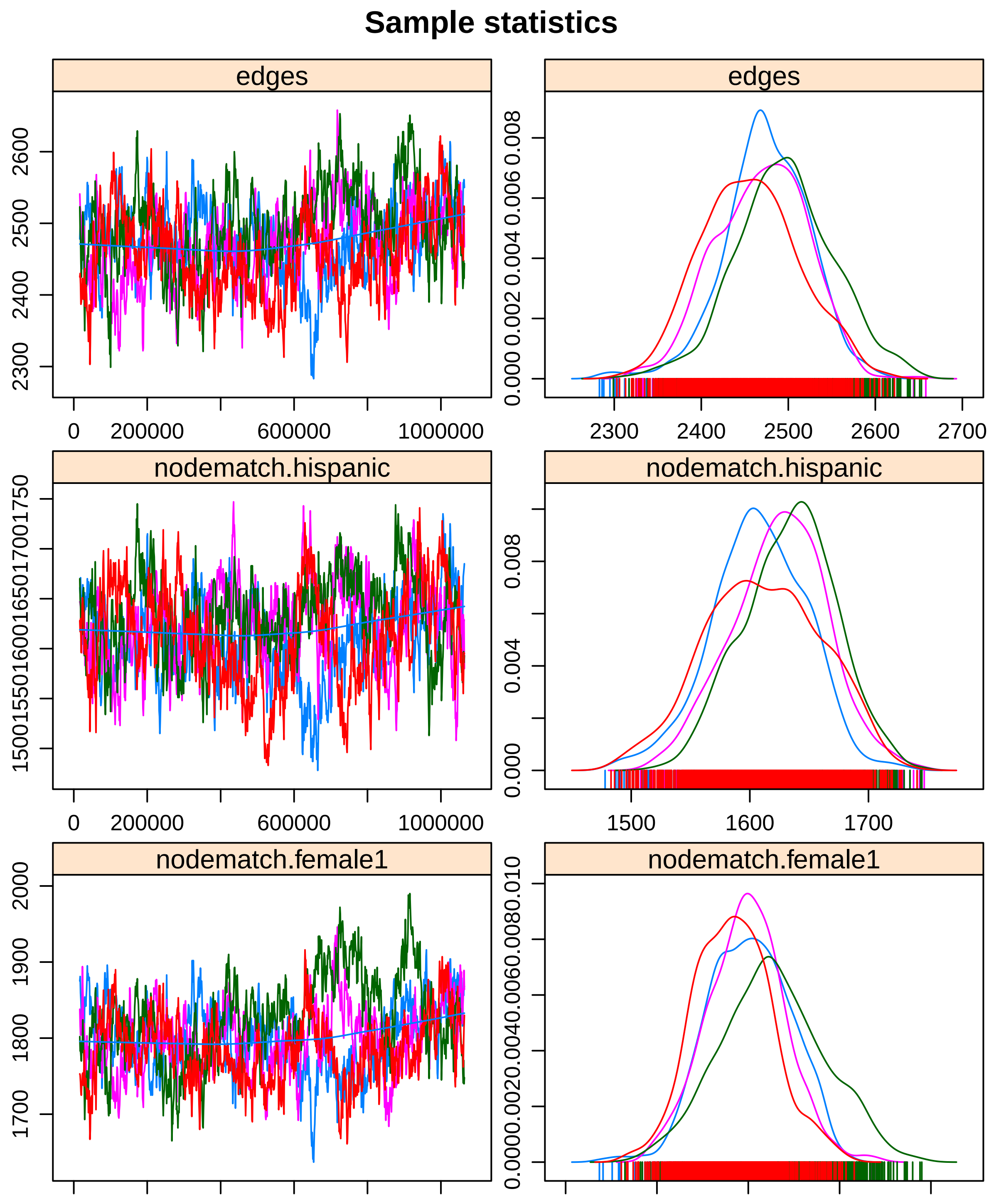


Figure 5.2: Trace and posterior distribution of sampled network statistics.

If we called the function `mcmc.diagnostics` this message appears at the end:

> MCMC diagnostics shown here are from the last round of simulation, prior to computation of final parameter estimates. Because the final estimates are refinements of

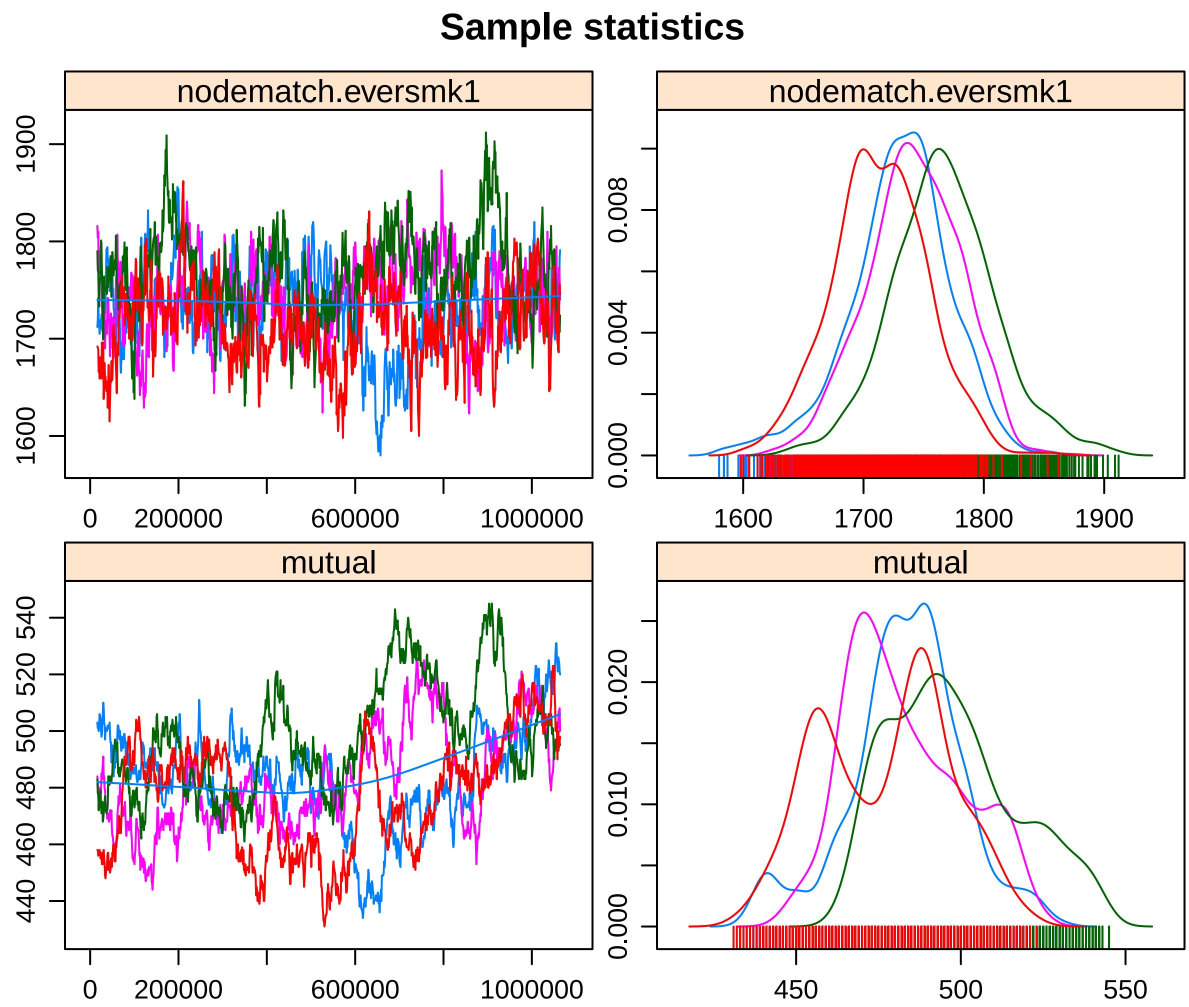


Figure 5.3: Trace and posterior distribution of sampled network statistics (cont'd).

those used for this simulation run, these diagnostics may understate model performance. To directly assess the performance of the final model on in-model statistics, please use the GOF command: gof(ergmFitObject, GOF=~model).

—`mcmc.diagnostics(ans0)`

Not that bad (although the `mutual` term could do better)![4] First, observe that in the plot we see 4 different lines, why is that? Well, since we were running in parallel using 4 cores the algorithm actually ran 4 different chains of the MCMC algorithm. An eyeball test is to see if all the chains moved at about the same place, if we have that we can start thinking about model convergence from the mcmc perspective.

Once we are sure to have reach convergence on the MCMC algorithm, we can start thinking about how well does our model predicts the observed network's proterties. Besides of the statistics that define our ERGM, the `gof` function's default behavior show GOF for:

a. In degree distribution,

[4] The statnet wiki website as a very nice example of (very) bad and good mcmc diagnostics plots here.

b. Out degree distribution,
c. Edge-wise shared partners, and
d. Geodesics

Let's take a look at it

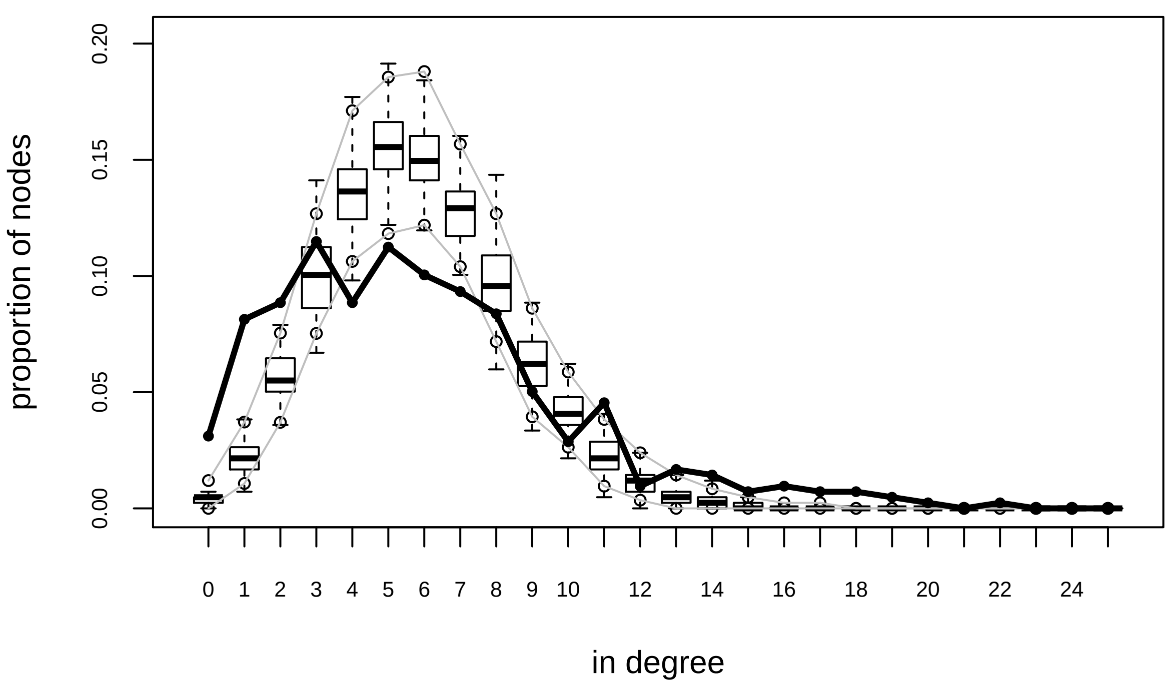
```r
# Computing and printing GOF estatistics
ans_gof <- gof(ans0)
ans_gof
```

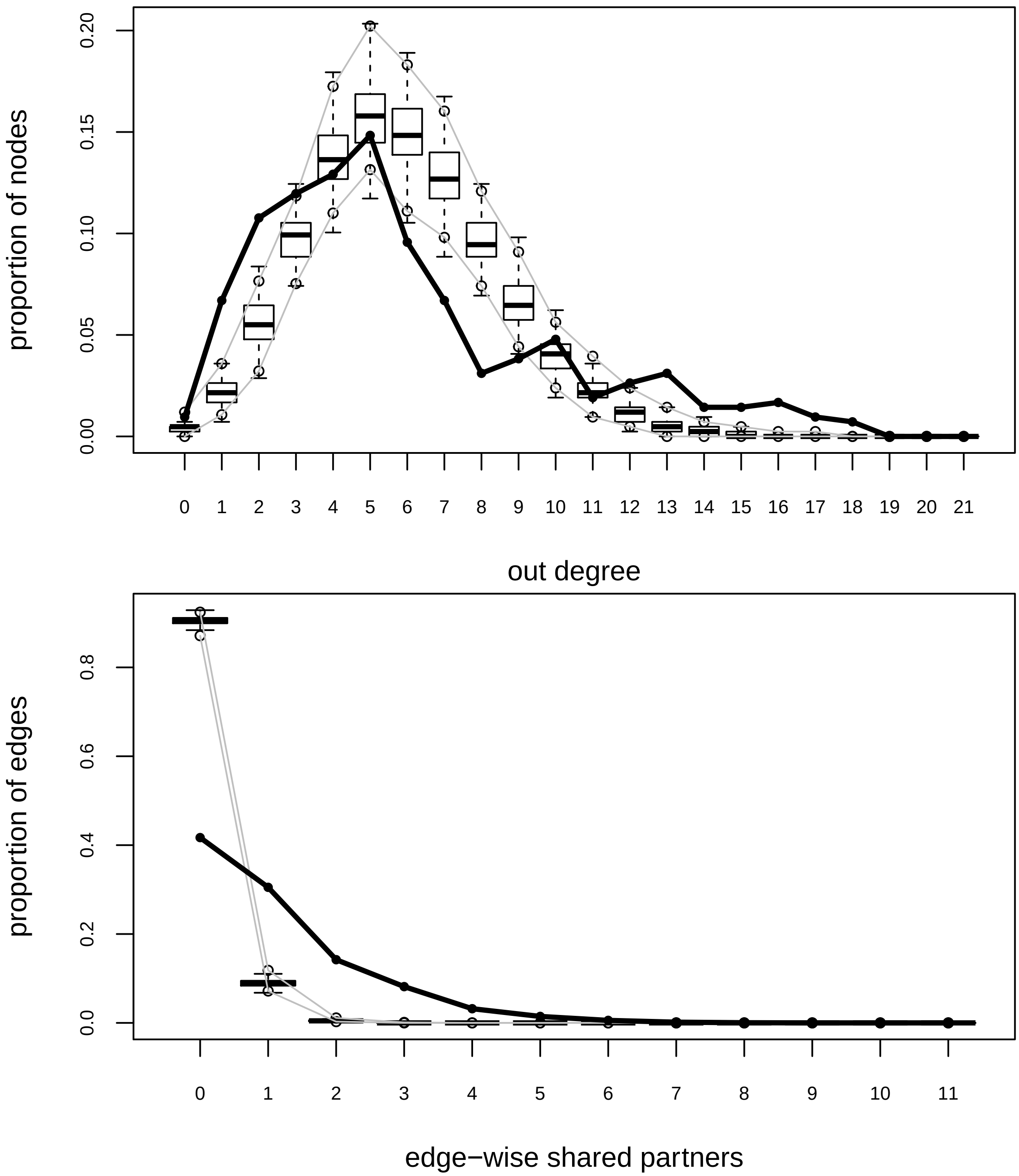
```
##
## Goodness-of-fit for in-degree
##
##    obs min  mean max MC p-value
## 0   13   0  1.89   8       0.00
## 1   34   3  9.04  18       0.00
## 2   37  11 23.63  33       0.00
## 3   48  28 41.83  59       0.44
## 4   37  41 56.87  75       0.00
## 5   47  44 64.71  84       0.04
## 6   42  39 63.33  85       0.02
## 7   39  42 53.78  74       0.00
## 8   35  25 40.58  60       0.50
## 9   21  14 26.19  43       0.38
## 10  12   9 17.37  26       0.16
## 11  19   2  9.53  17       0.00
## 12   4   0  4.93  11       0.90
## 13   7   0  2.35   7       0.04
## 14   6   0  1.27   5       0.00
## 15   3   0  0.44   3       0.02
## 16   4   0  0.21   2       0.00
## 17   3   0  0.05   1       0.00
## 18   3   0  0.00   0       0.00
## 19   2   0  0.00   0       0.00
## 20   1   0  0.00   0       0.00
## 22   1   0  0.00   0       0.00
##
## Goodness-of-fit for out-degree
##
##    obs min  mean max MC p-value
## 0    4   0  1.85   5       0.20
## 1   28   3  8.99  15       0.00
## 2   45  12 23.25  35       0.00
## 3   50  24 40.87  52       0.06
## 4   54  42 57.89  76       0.68
```

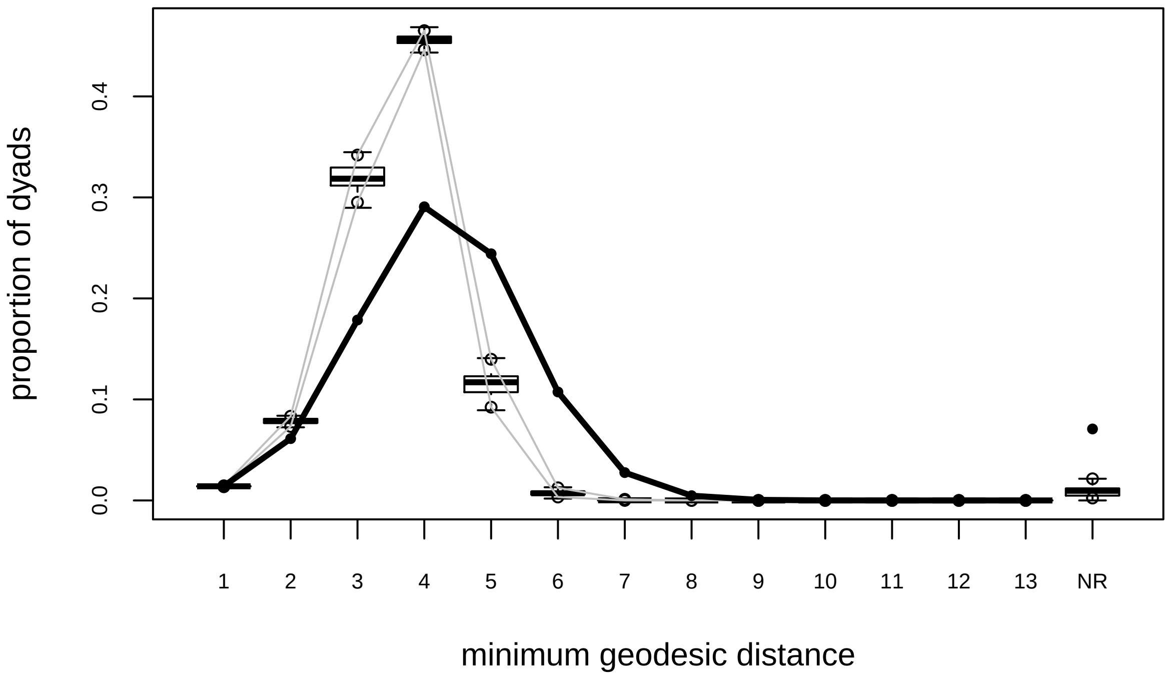
```
## 5   62  49 66.04  85       0.70
## 6   40  41 62.23  79       0.00
## 7   28  37 54.08  70       0.00
## 8   13  29 40.05  52       0.00
## 9   16  17 27.65  41       0.00
## 10  20   8 16.72  30       0.46
## 11   8   2  9.30  19       0.76
## 12  11   1  4.98  11       0.04
## 13  13   0  2.38   7       0.00
## 14   6   0  0.97   4       0.00
## 15   6   0  0.50   3       0.00
## 16   7   0  0.17   1       0.00
## 17   4   0  0.06   1       0.00
## 18   3   0  0.01   1       0.00
## 19   0   0  0.01   1       1.00
##
## Goodness-of-fit for edgewise shared partner
##
##       obs  min    mean  max MC p-value
## esp0 1032 2012 2210.11 2303          0
## esp1  755  156  222.10  441          0
## esp2  352    4   13.42   93          0
## esp3  202    0    0.77   19          0
## esp4   79    0    0.04    3          0
## esp5   36    0    0.00    0          0
## esp6   14    0    0.00    0          0
## esp7    4    0    0.00    0          0
## esp8    1    0    0.00    0          0
##
## Goodness-of-fit for minimum geodesic distance
##
##       obs   min     mean   max MC p-value
## 1    2475  2301  2446.44  2568       0.56
## 2   10672 12062 13688.54 14617       0.00
## 3   31134 48722 55636.04 60092       0.00
## 4   50673 77284 79447.41 81661       0.00
## 5   42563 14452 20165.40 26886       0.00
## 6   18719   325  1274.88  2453       0.00
## 7    4808     1    51.78   361       0.00
## 8     822     0     2.13   102       0.00
## 9     100     0     0.06     4       0.00
## 10      7     0     0.01     1       0.00
## Inf 12333     0  1593.31  4558       0.00
```

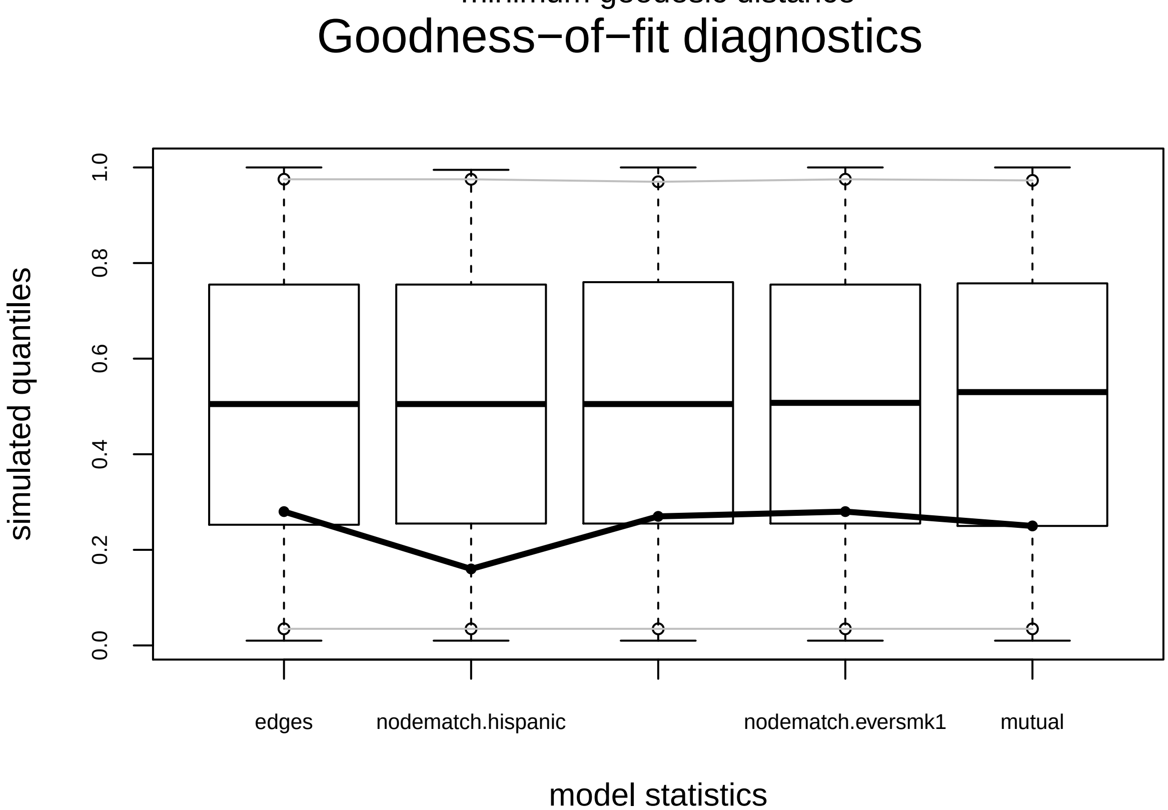
```
##
## Goodness-of-fit for model statistics
##
##                      obs  min    mean  max MC p-value
## edges               2475 2301 2446.44 2568       0.56
## nodematch.hispanic 1615 1499 1578.58 1662       0.32
## nodematch.female1  1814 1690 1791.16 1883       0.54
## nodematch.eversmk1 1738 1595 1716.19 1834       0.56
## mutual               486  436  475.48  504       0.50
```

```
# Plotting GOF statistics
plot(ans_gof)
```




proportion of nodes
0.00
0.05
0.10
0.15
0.20
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
out degree
proportion of edges
0.0
0.2
0.4
0.6
0.8
0 1 2 3 4 5 6 7 8 9 10 11
edge−wise shared partners





Try the following configuration instead

```
ans0_bis <- ergm(
  network_111 ~
    edges +
    nodematch("hispanic") +
```

```
    nodematch("female1") +
    mutual +
    esp(0:3) +
    idegree(0:10)
    ,
  constraints = ~bd(maxout = 19),
  control = control.ergm(
    seed         = 1,
    MCMLE.maxit = 15,
    parallel     = 4,
    CD.maxit     = 15,
    MCMC.samplesize = 2048*4,
    MCMC.burnin = 30000,
    MCMC.interval = 2048*4
    )
  )
```

Increase the sample size so the curves are more smooth, longer intervals (thinning) so we reduce the autocorrelation, larger burin. All this together to improve the Gelman test statistic. We also added idegree from 0 to 10, and esp from 0 to 3 to explicitly match those statistics in our model.

```
knitr::include_graphics("awful-chains.png")
```

5.5 More on MCMC convergence

For more on this issue, I recommend reviewing chapter 1 and chapter 6 from the Handbook of MCMC (Brooks et al. 2011). Both chapters are free to download from the book's website.

For GOF take a look at section 6 of ERGM 2016 Sunbelt tutorial, and for a more technical review you can take a look at (David R. Hunter, Goodreau, and Handcock 2008).

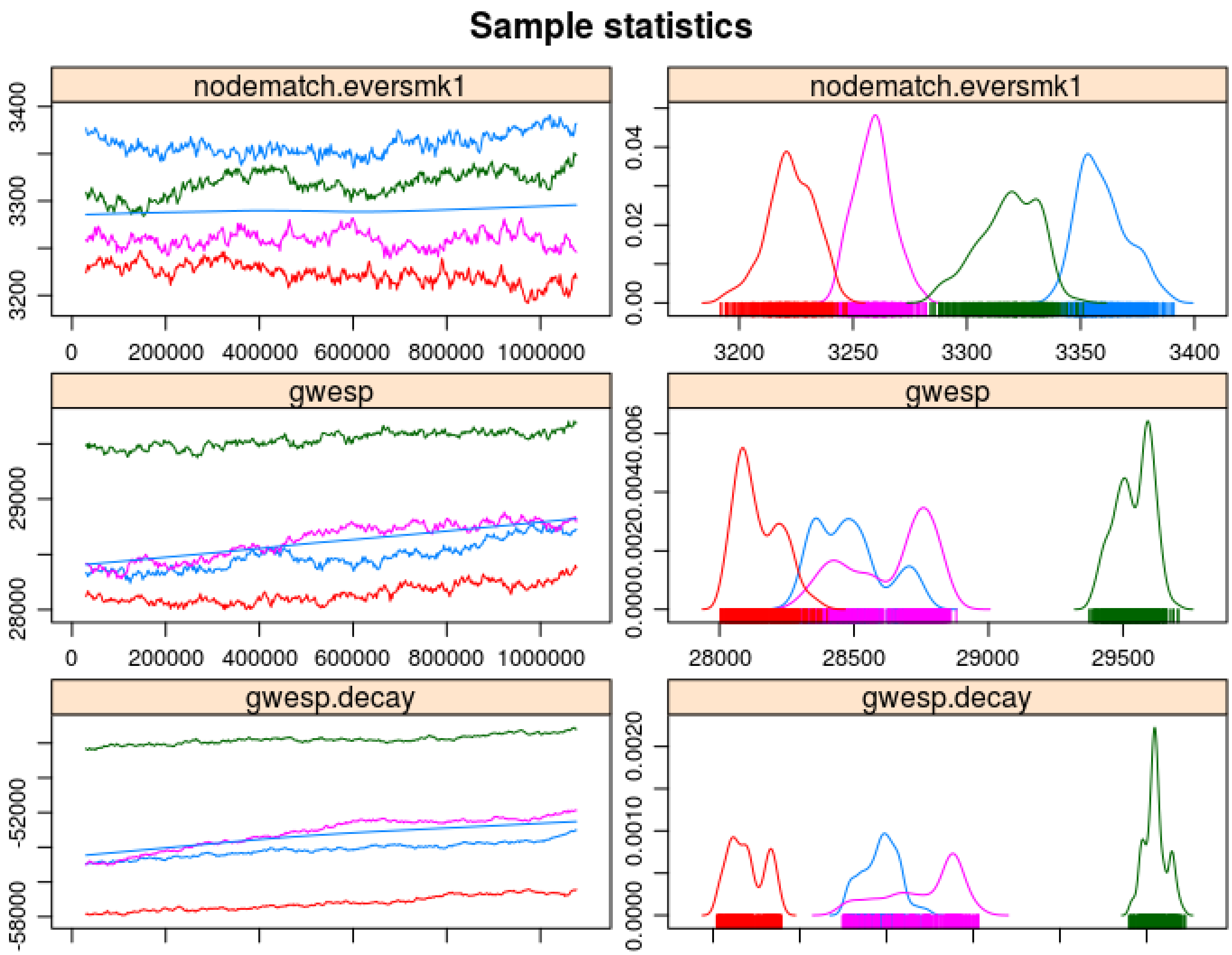


Figure 5.4: An example of a terrible ERGM (no convergence at all). Also, a good example of why running multiple chains can be useful

Chapter 6

(Separable) Temporal Exponential Family Random Graph Models

This tutorial is great! https://statnet.org/trac/raw-attachment/wiki/Sunbelt2016/tergm_tutorial.pdf

Chapter 7

Stochastic Actor Oriented Models

Stochastic Actor Oriented Models (SOAM), also known as Siena models were introduced by CITATION NEEDED.

As a difference from ERGMs, Siena models look at the data generating process from the individuals' point of view. Based on McFadden's ideas of probabilistic choice, the model is founded in the following equation

$$U_i(x) - U_i(x') \sim \text{Extreame Value Distribution}$$

In other words, individuals choose between states x and x' in a probabilistic way (with some noise),

$$\frac{\exp\left\{f_i^Z(\beta^z, x, z)\right\}}{\sum_{Z' \in \mathcal{C}} \exp\left\{f_i^Z(\beta, x, z')\right\}}$$

snijders_(sociological methodology 2001)

Ripley et al. (2011)

Chapter 8

Hypothesis testing in networks

Overall, there are many ways in which we can see hypothesis testing within the networks context:

1. **Comparing two or more networks**, e.g., we want to see if the density of two networks are *equal*.
2. **Prevalence of a motif/pattern**, e.g., check whether the observed number of transitive triads is different from that expected as of by chance.
3. **Multivariate using ERGMs**, e.g., jointly test whether homophily and two stars are the motifs that drive network structure.

The latter we already review in the ERGM chapter. In this part, we will look at types one and two; both using non-parametric methods.

8.1 Comparing networks

Imagine that we have two graphs, $(G_1, G_2) \in \mathcal{G}$, and we would like to assess whether a given statistic $s(\cdot)$, e.g., density, is equal in both of them. Formally, we would like to asses whether $H_0: s(G_1) - s(G_2) = k$ vs $H_a: s(G_1) - s(G_2) \neq k$.

As usual, the true distribution of $s(\cdot)$ is unknown, thus, one approach that we could use is a non-parametric bootstrap test.

8.1.1 Network bootstrap

The non parametric bootstrap and jackknife methods for social networks were introduced by (T. A. B. Snijders and Borgatti 1999). The method itself is used to generate standard errors for network level statistics. Both methods are implemented in the R package `netdiffuseR`.

8.1.2 When the statistic is normal

When the we deal with things that are normally distributed, e.g., sample means like density[1], we can make use of the Student's distribution for making inference. In particular, we can use Bootstrap/Jackknife to approximate the standard errors of the statistic for each network:

1. Since $s(G_i) \sim \mathrm{N}(\mu_i, \sigma_i^2/m_i)$ for $i \in \{1, 2\}$, in the case of the density, $m_i = n_i * (n_i - 1)$. The statistic is then:

$$s(G_1) - s(G_0) \sim \mathrm{N}(\mu_1 - \mu_0, \sigma_1^2/m_1 + \sigma_1^2/m_2)$$

   Thus

$$\frac{s(G_1) - s(G_0) - \mu_1 + \mu_2}{\sqrt{\sigma_1^2/m_1 + \sigma_1^2/m_2}} \sim t_{m_1+m_2-2}$$

   But, if we are testing $H_0 : \mu_1 - \mu_2 = k$, then, under the null

$$\frac{s(G_1) - s(G_0) - k}{\sqrt{\sigma_1^2/m_1 + \sigma_1^2/m_2}} \sim t_{m_1+m_2-2}$$

   Where We now proceede to approximate the variances.

2. Using the *plugin principle* (Efron and Tibshirani 1994), we can approximate the variances using Bootstrap/Jackknife, i.e., compute $\hat{\sigma}_1^2 \approx \sigma_1^2/m_1$ and $\hat{\sigma}_2^2 \approx \sigma_2^2/m_2$. Using netdiffuseR

```r
library(netdiffuseR)

# Obtain a 100 replicates
sg1 <- bootnet(g1, function(i) sum(i)/(nnodes(i) * (nnodes(i) - 1)), R = 100)
sg2 <- bootnet(g2, function(i) sum(i)/(nnodes(i) * (nnodes(i) - 1)), R = 100)

# Retrieving the variances
hat_sigma1 <- sg1$var_t
hat_sigma2 <- sg2$var_t

# And the actual values
sg1 <- sg1$t0
sg2 <- sg2$t0
```

3. With the approximates in hand, we can then use the the "t-test table" to retrieve the corresponding value, in R:

[1] Density is indeed a sample mean as we are, in principle computing the average of a sequence of Bernoulli variables. Formally: $\mathrm{density}(G) = \frac{1}{n(n-1)} \sum_{ij} A_{ij}$.

```r
# Building the statistic
tstat <- (sg1 - sg2 - k)/(sqrt(hat_sigma1 + hat_sigma2))

# Computing the pvalue
m1 <- nnodes(g1)*(nnodes(g1) - 1)
m2 <- nnodes(g2)*(nnodes(g2) - 1)
pt(tstat, df = m1 + m2 - 2)
```

8.1.3 When the statistic is NOT normal

In the case that the statistic is not normally distributed, we cannot use the t-statistic any longer. Nevertheless, the Bootstrap can come to help. While in general it is better to use distributions of pivot statistics (see (Efron and Tibshirani 1994)), we can still leverage the power of this method to make inferences. For this example, $s(\cdot)$ will be the range of the threshold in a diffusion graph.

As before, imagine that we are dealing with an statistic $s(\cdot)$ for two different networks, and we would like to asses whether we can reject H_0 or fail to reject it. The procedure is very similar:

1. One approach that we can test is whether $k \in \text{ConfInt}(s(G_1) - s(G_2))$. Building confidence intervals with bootstrap could be more intuitive.

2. Like before, we use bootstrap to generate a distribution of $s(G_1)$ and $s(G_2)$, in R:

```r
# Obtain a 1000 replicates
sg1 <- bootnet(g1, function(i) range(threshold(i)), R = 1000)
sg2 <- bootnet(g2, function(i) range(threshold(i)), R = 1000)

# Retrieving the distributions
sg1 <- sg1$boot$t
sg2 <- sg2$boot$t

# Define the statistic
sdiff <- sg1 - sg2
```

3. Once we have `sdiff`, we can proceed and compute the, for example, 95% confidence interval, and evaluate whether k falls within. In R:

```r
diff_ci <- quantile(sdiff, probs = c(0.025, .975))
```

This corresponds to what Efron and Tibshirani call “percentile interval.” This is easy to compute, but a better approach is using the “BCa” method, “Bias Corrected and Accelerated.” (TBD)

8.2 Examples

8.2.1 Average of node-level stats

Supposed that we would like to compare something like average indegree. In particular, for both networks, G_1 and G_2, we compute the average indegree per node:

$$s(G_1) = \text{AvgIndeg}(G_1) = \frac{1}{n}\sum_i \sum_{j\neq i} A^1_{ji}$$

where A^1_{ji} equals one if vertex j sends a tie to i. In this case, since we are looking at an average, we have that $\text{AvgIndeg}(G_1) \sim N(\mu_1, \sigma_1^2/n)$. Thus, taking advantage of the normality of the statistic, we can build a test statistic as follows:

$$\frac{s(G_1) - s(G_2) - k}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}} \sim t_{n_1+n_2-2}$$

Where $\hat{\sigma}_i$ is the bootstrap standard error, and $k = 0$ when we are testing equality. This distributes t with $n_1 + n_2 - 2$ degrees of freedom. As a difference from the previous example using density, the degrees of freedom for this test are less as, instead of having an average across all entries of the adjacency matrix, we have an average across all vertices.

Appendix A

Datasets

A.1 SNS data

A.1.1 About the data

- This data is part of the NIH Challenge grant # RC 1RC1AA019239 "Social Networks and Networking That Puts Adolescents at High Risk."
- In general terms, the SNS's goal was(is) "Understand the network effects on risk behaviors such as smoking initiation and substance use."

A.1.2 Variables

The data has a *wide* structure, which means that there is one row per individual, and that dynamic attributes are represented as one column per time.

- `photoid` Photo id at the school level (can be repeated across schools).
- `school` School id.
- `hispanic` Indicator variable that equals 1 if the indivual ever reported himself as hispanic.
- `female1`, ..., `female4` Indicator variable that equals 1 if the individual reported to be female at the particular wave.
- `grades1`,..., `grades4` Academic grades by wave. Values from 1 to 5, with 5 been the best.
- `eversmk1`, ..., `eversmk4` Indicator variable of ever smoking by wave. A one indicated that the individual had smoked at the time of the survey.
- `everdrk1`, ..., `everdrk4` Indicator variable of ever drinking by wave. A one indicated that the individual had drink at the time of the survey.
- `home1`, ..., `home4` Factor variable for home status by wave. A one indicates home ownership, a 2 rent, and a 3 a "I don't know."

During the survey, participants were asked to name up to 19 of their school friends:

- `sch_friend11`, ..., `sch_friend119` School friends nominations (19 in total) for wave 1. The codes are mapped to the variable `photoid`.
- `sch_friend21`, ..., `sch_friend219` School friends nominations (19 in total) for wave 2. The codes are mapped to the variable `photoid`.
- `sch_friend31`, ..., `sch_friend319` School friends nominations (19 in total) for wave 3. The codes are mapped to the variable `photoid`.
- `sch_friend41`, ..., `sch_friend419` School friends nominations (19 in total) for wave 4. The codes are mapped to the variable `photoid`.

References

Admiraal, Ryan, and Mark S Handcock. 2006. "Sequential Importance Sampling for Bipartite Graphs with Applications to Likelihood-Based Inference." Department of Statistics, University of Washington.

Bache, Stefan Milton, and Hadley Wickham. 2014. *Magrittr: A Forward-Pipe Operator for r*. https://CRAN.R-project.org/package=magrittr.

Bojanowski, Michal. 2015. *Intergraph: Coercion Routines for Network Data Objects*. http://mbojan.github.io/intergraph.

Brooks, Steve, Andrew Gelman, Galin Jones, and Xiao-Li Meng. 2011. *Handbook of Markov Chain Monte Carlo*. CRC press.

Csardi, Gabor, and Tamas Nepusz. 2006. "The Igraph Software Package for Complex Network Research." *InterJournal* Complex Systems: 1695. http://igraph.org.

Efron, Bradley, and Robert J Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC press.

Geyer, Charles J., and Elizabeth A. Thompson. 1992. "Constrained Monte Carlo Maximum Likelihood for Dependent Data." *Journal of the Royal Statistical Society. Series B (Methodological)* 54 (3): 657–99. http://www.jstor.org/stable/2345852.

Handcock, Mark S., David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky, Skye Bender-deMoll, and Martina Morris. 2016. *Statnet: Software Tools for the Statistical Analysis of Network Data*. The Statnet Project (http://www.statnet.org). CRAN.R-project.org/package=statnet.

Handcock, Mark S., David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky, and Martina Morris. 2017. *Ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project (http://www.statnet.org). https://CRAN.R-project.org/package=ergm.

Hunter, David R., Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. 2008. "ergm : A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks." *Journal of Statistical Software* 24 (3). https://doi.org/10.18637/jss.v024.i03.

Hunter, David R, Steven M Goodreau, and Mark S Handcock. 2008. "Goodness of Fit of Social Network Models." *Journal of the American Statistical Association* 103 (481): 248–58.

https://doi.org/10.1198/016214507000000446.

Lazega, Emmanuel, and Tom AB Snijders. 2015. *Multilevel Network Analysis for the Social Sciences: Theory, Methods and Applications*. Vol. 12. Springer.

Leifeld, Philip. 2013. "texreg: Conversion of Statistical Model Output in R to LaTeX and HTML Tables." *Journal of Statistical Software* 55 (8): 1–24. http://www.jstatsoft.org/v55/i08/.

Lusher, Dean, Johan Koskinen, and Garry Robins. 2012. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge University Press.

Matloff, Norman. 2011. *The Art of r Programming: A Tour of Statistical Software Design*. No Starch Press.

Morris, Martina, Mark Handcock, and David Hunter. 2008. "Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects." *Journal of Statistical Software, Articles* 24 (4): 1–24. https://doi.org/10.18637/jss.v024.i04.

Plummer, Martyn, Nicky Best, Kate Cowles, and Karen Vines. 2006. "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News* 6 (1): 7–11. https://journal.r-project.org/archive/.

R Core Team. 2017a. *Foreign: Read Data Stored by 'Minitab', 's', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ...* https://CRAN.R-project.org/package=foreign.

———. 2017b. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ripley, Ruth M., Tom AB Snijders, Paulina Preciado, and Others. 2011. "Manual for RSIENA." *University of Oxford: Department of Statistics, Nuffield College*, no. 2007. https://www.uni-due.de/hummell/sna/R/RSiena_Manual.pdf.

Sarkar, Deepayan, and Felix Andrews. 2016. *latticeExtra: Extra Graphical Utilities Based on Lattice*. https://CRAN.R-project.org/package=latticeExtra.

Snijders, Tom A B, and Stephen P Borgatti. 1999. "Non-Parametric Standard Errors and Tests for Network Statistics." *Connections* 22 (2): 1–10. https://insna.org/PDF/Connections/v22/1999_I-2_61-70.pdf.

Snijders, Tom A B, Gerhard G. van de Bunt, and Christian E G Steglich. 2010. "Introduction to stochastic actor-based models for network dynamics." *Social Networks* 32 (1): 44–60. https://doi.org/10.1016/j.socnet.2009.02.004.

Snijders, Tom AB. 2002. "Markov Chain Monte Carlo Estimation of Exponential Random Graph Models." *Journal of Social Structure* 3.

Ushey, Kevin, Jim Hester, and Robert Krzyzanowski. 2017. *Rex: Friendly Regular Expressions*. https://CRAN.R-project.org/package=rex.

Wang, Peng, Ken Sharpe, Garry L. Robins, and Philippa E. Pattison. 2009. "Exponential Random Graph (p*) Models for Affiliation Networks." *Social Networks* 31 (1): 12–25. https://doi.org/https://doi.org/10.1016/j.socnet.2008.08.002.

Wickham, Hadley. 2017. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, and Jennifer Bryan. 2017. *Readxl: Read Excel Files*. https://CRAN.R-project.org/package=readxl.

Wickham, Hadley, Romain Francois, Lionel Henry, and Kirill Müller. 2017. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Lionel Henry. 2017. *Tidyr: Easily Tidy Data with 'Spread()' and 'Gather()' Functions*. https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley, Jim Hester, and Romain Francois. 2017. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.