ETC1010: Introduction to Data Analysis

Week 9, part A

Networks and Graphs

Lecturer: *Nicholas Tierney*

Department of Econometrics and Business Statistics

nicholas.tierney@monash.edu

May 2020

Announcements

- Project deadlines:
  - **Deadline 2 (22nd May)** : Team members and team name, data description.
  - **Deadline 3 (29th May)** : Electronic copy of your data, and a page of data description, and cleaning done, or needing to be done.
- Practical exam.

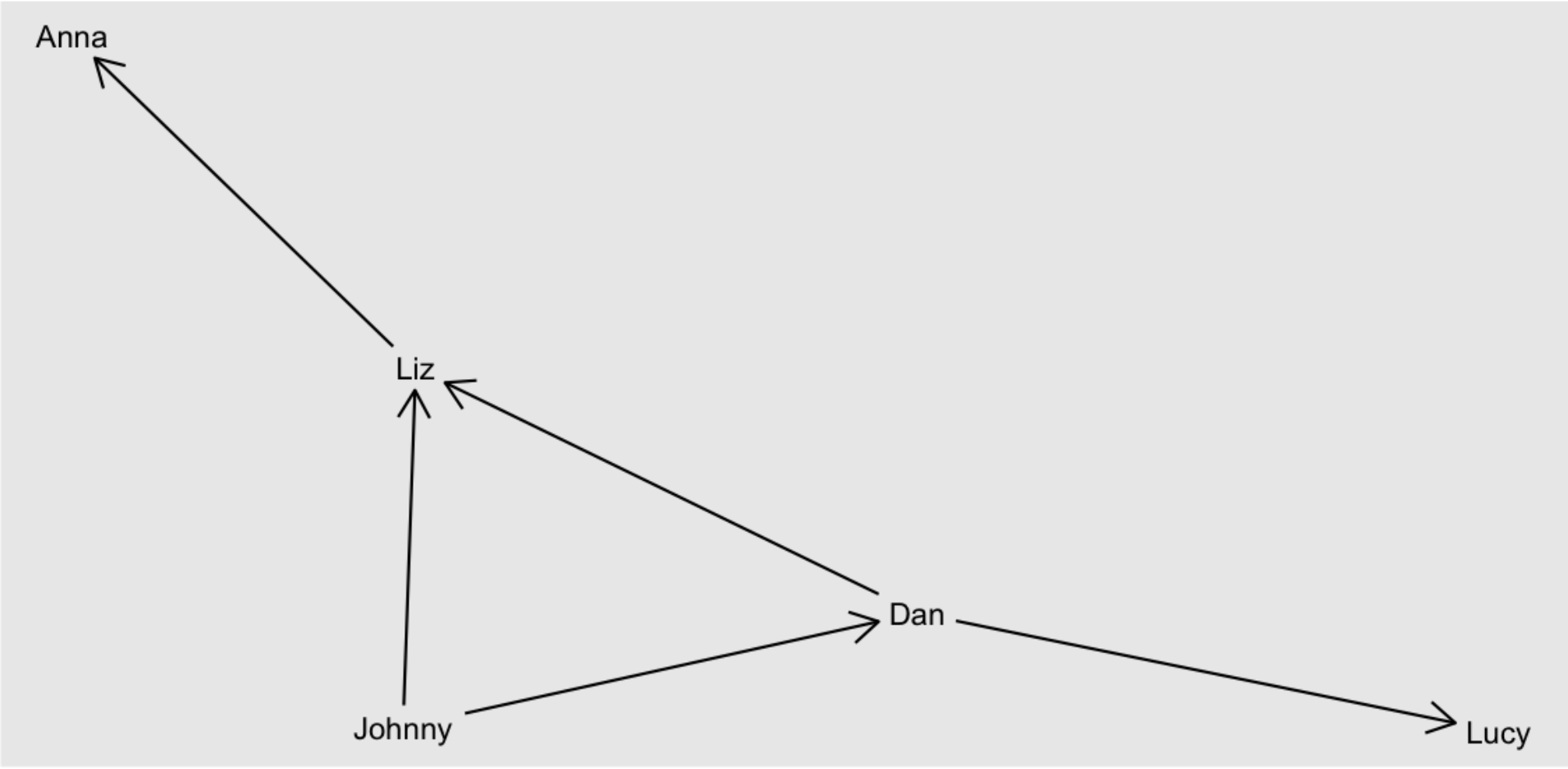
recap: Last week on tidy text data

Network analysis

A description of phone calls

- Johnny --> Liz
- Liz --> Anna
- Johnny -- > Dan
- Dan --> Liz
- Dan --> Lucy

As a graph



And as an association matrix

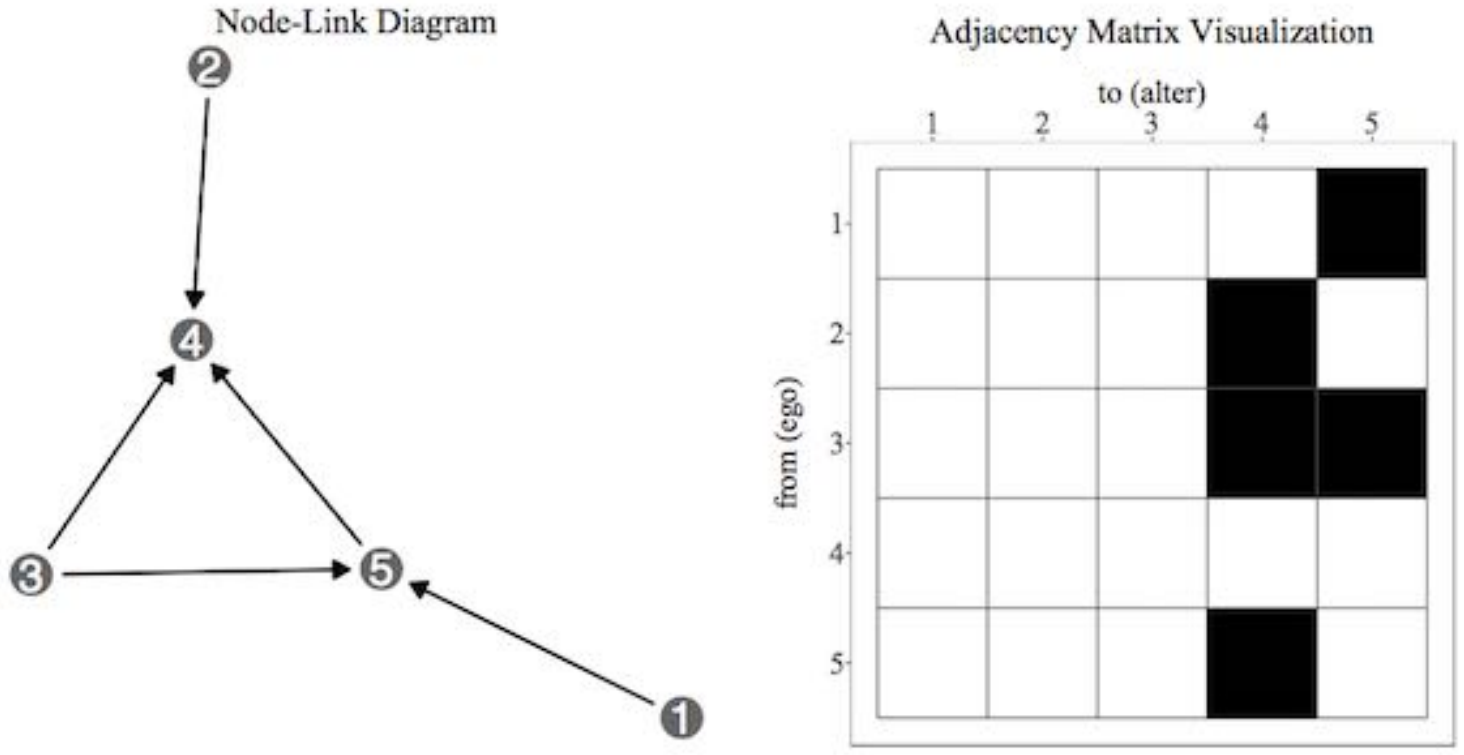
[DEMO]

Why care about these relationships?

- **Telephone exchanges**: Nodes are the phone numbers. Edges would indicate a call was made betwen two numbers.
- **Book or movie plots**: Nodes are the characters. Edges would indicate whether they appear together in a scene, or chapter. If they speak to each other, various ways we might measure the association.
- **Social media**: nodes would be the people who post on facebook, including comments. Edges would measure who comments on who's posts.

Drawing these relationships out:

One way to describe these relationships is to provide association matrix between many objects.



(Image created by Sam Tyner.)

Example: Madmen

Source: wikicommons

Generate a network view

- Create a layout (in 2D) which places nodes which are most related close,
- Plot the nodes as points, connect the appropriate lines
- Overlaying other aspects, e.g. gender

introducing madmen data

```
glimpse(madmen)
## List of 2
##  $ edges   :'data.frame':    39 obs. of  2 variables:
##   ..$ Name1: Factor w/ 9 levels "Betty Draper",..: 1 1 2 2 2 2 2 2 2 2 ...
##   ..$ Name2: Factor w/ 39 levels "Abe Drexler",..: 15 31 2 4 5 6 8 9 11 21 ...
##  $ vertices:'data.frame':    45 obs. of  2 variables:
##   ..$ label : Factor w/ 45 levels "Abe Drexler",..: 5 9 16 23 26 32 33 38 39 17 ...
##   ..$ Gender: Factor w/ 2 levels "female","male": 1 2 2 1 2 1 2 2 2 2 ...
```

Nodes and edges?

Netword data can be thought of as two related tables, **nodes** and **edges**:

- **nodes** are connection points
- **edges** are the connections between points

Example: Mad Men. (Nodes = characters from the series)

```
madmen_nodes
## # A tibble: 45 x 2
##    label          gender
##    <chr>          <chr>
##  1 Betty Draper   female
##  2 Don Draper     male
##  3 Harry Crane    male
##  4 Joan Holloway  female
##  5 Lane Pryce     male
##  6 Peggy Olson    female
##  7 Pete Campbell  male
##  8 Roger Sterling male
##  9 Sal Romano     male
## 10 Henry Francis  male
## # … with 35 more rows
```

Example: Mad Men. (Edges = how they are associated)

```
madmen_edges
```

```
## # A tibble: 39 x 2
##    Name1         Name2
##    <chr>         <chr>
##  1 Betty Draper Henry Francis
##  2 Betty Draper Random guy
##  3 Don Draper   Allison
##  4 Don Draper   Bethany Van Nuys
##  5 Don Draper   Betty Draper
##  6 Don Draper   Bobbie Barrett
##  7 Don Draper   Candace
##  8 Don Draper   Doris
##  9 Don Draper   Faye Miller
## 10 Don Draper   Joy
## # … with 29 more rows
```

Let's get the madmen data into the right shape

```
madmen_edges %>%
  rename(from_id = Name1, to_id = Name2)
## # A tibble: 39 x 2
##    from_id      to_id
##    <chr>        <chr>
##  1 Betty Draper Henry Francis
##  2 Betty Draper Random guy
##  3 Don Draper   Allison
##  4 Don Draper   Bethany Van Nuys
##  5 Don Draper   Betty Draper
##  6 Don Draper   Bobbie Barrett
##  7 Don Draper   Candace
##  8 Don Draper   Doris
##  9 Don Draper   Faye Miller
## 10 Don Draper   Joy
## # … with 29 more rows
```

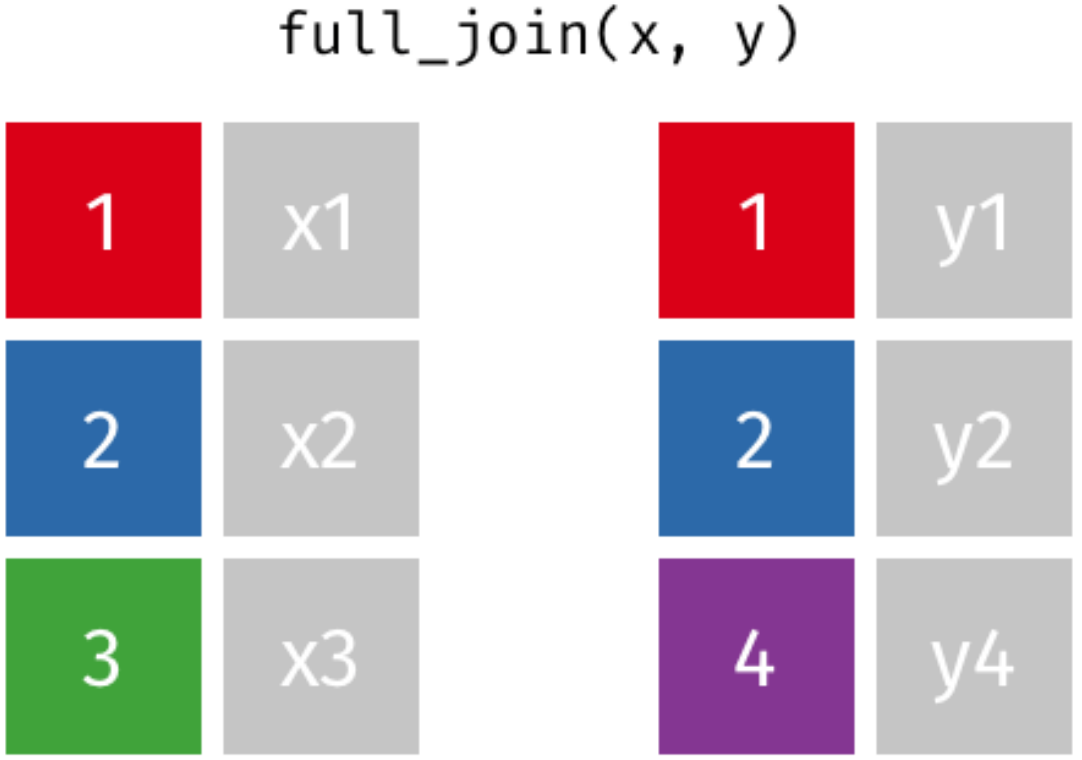
Let's get the madmen data into the right shape

```
madmen_net <- madmen_edges %>%
  rename(from_id = Name1, to_id = Name2) %>%
  full_join(madmen_nodes,
            by = c("from_id" = "label"))
```

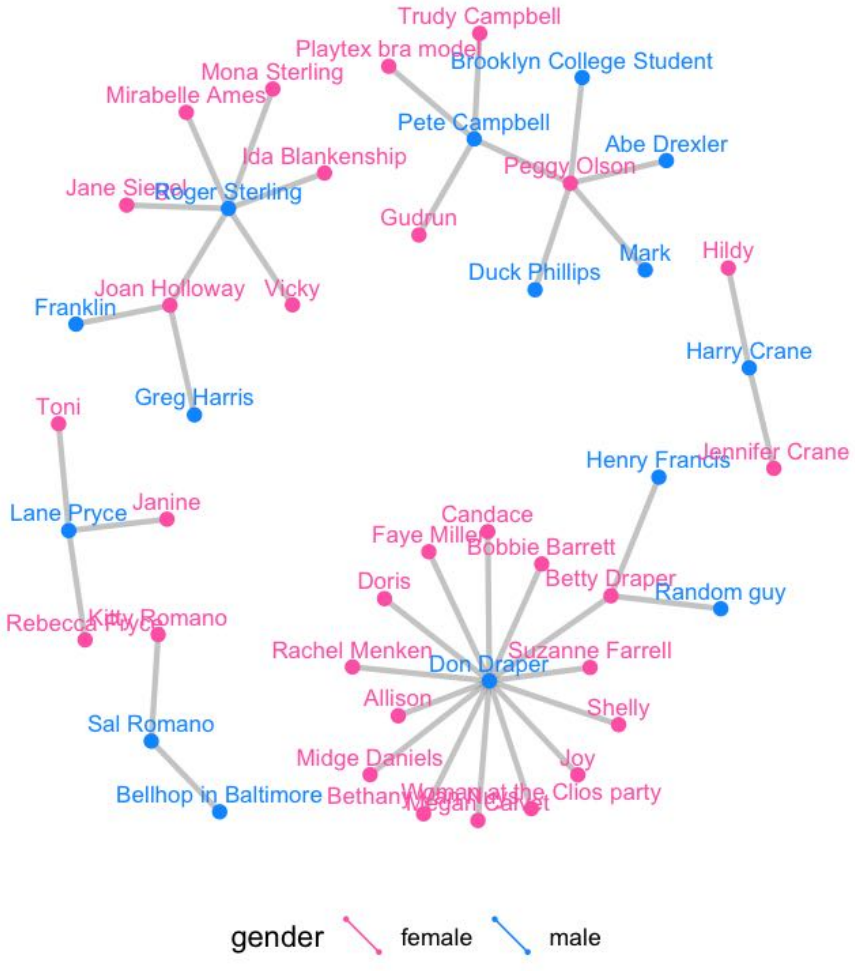
```
madmen_net
```

```
## # A tibble: 75 x 3
##    from_id      to_id            gender
##    <chr>        <chr>            <chr>
##  1 Betty Draper Henry Francis    female
##  2 Betty Draper Random guy       female
##  3 Don Draper   Allison          male
##  4 Don Draper   Bethany Van Nuys male
##  5 Don Draper   Betty Draper     male
##  6 Don Draper   Bobbie Barrett   male
##  7 Don Draper   Candace          male
##  8 Don Draper   Doris            male
##  9 Don Draper   Faye Miller      male
## 10 Don Draper   Joy              male
## # … with 65 more rows
```

Full join?



Plotting the data with geomnet



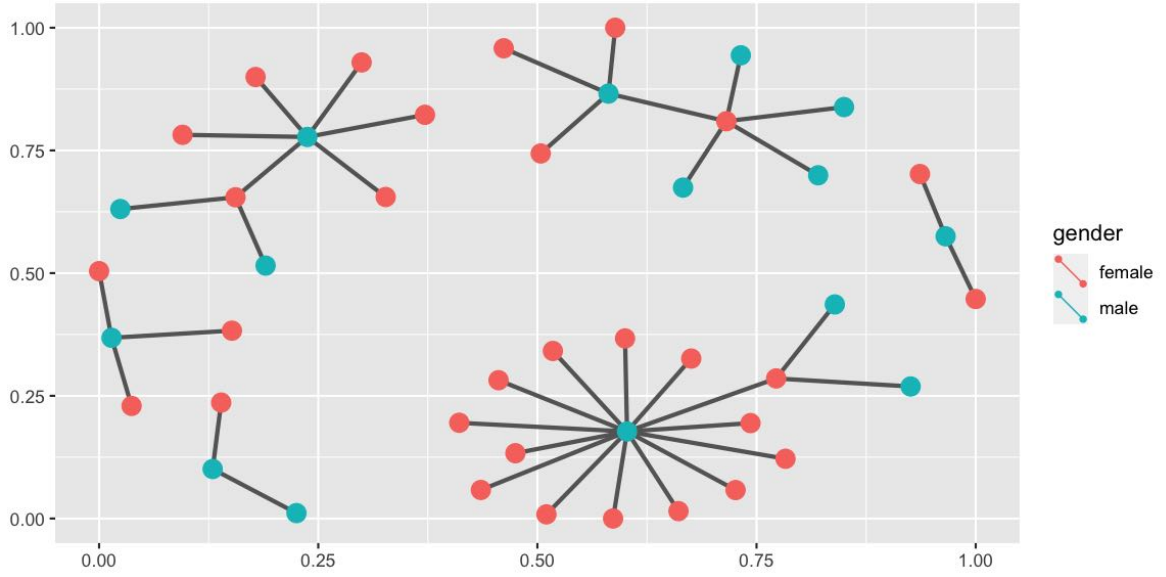
Aside: Installing `geomnet`

This is the code you will need to use to install it:

```
install.packages("remotes")
library(remotes)
install_github("sctyner/geomnet")
```

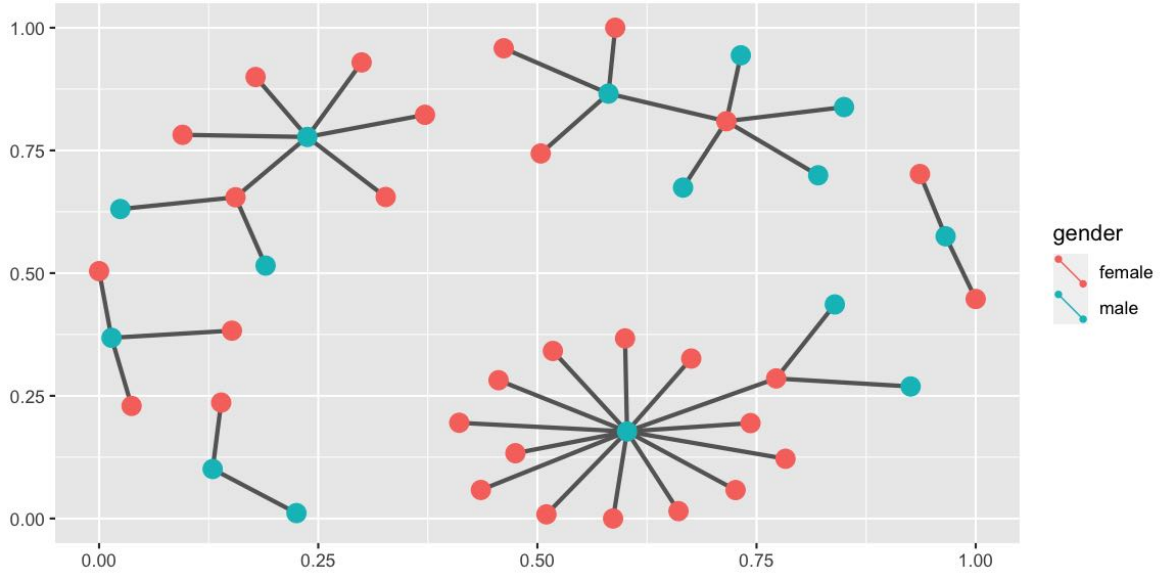
How to plot

```
set.seed(5556677)
ggplot(data = madmen_net,
       aes(from_id = from_id,
           to_id = to_id)) +
  geom_net(aes(colour = gender))
```



How to plot: specify the layout algorithm

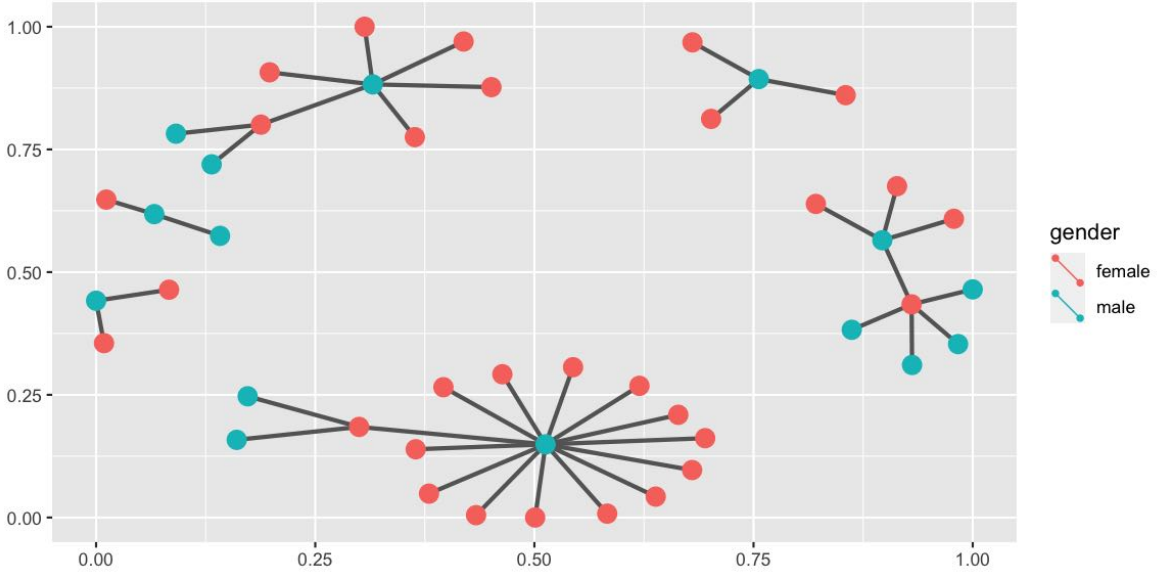
```
set.seed(5556677)
ggplot(data = madmen_net,
       aes(from_id = from_id,
           to_id = to_id)) +
  geom_net(aes(colour = gender),
           layout.alg = "kamadak
```



How to plot: Try different layout algorithms

Follow links in ?geom_net for more examples:

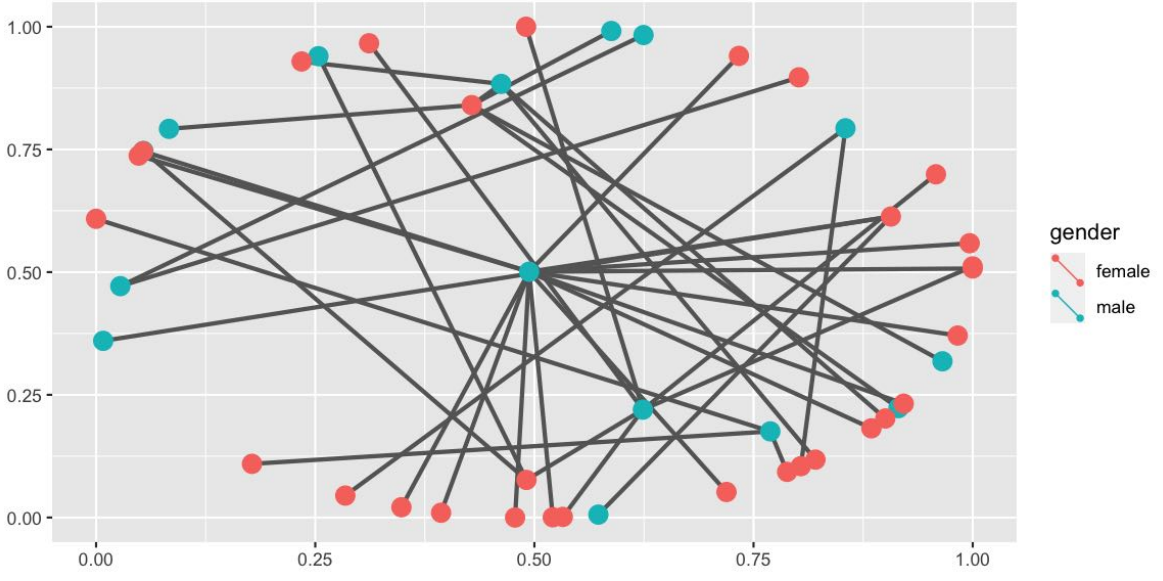
```
set.seed(5556677)
ggplot(data = madmen_net,
       aes(from_id = from_id,
           to_id = to_id)) +
  geom_net(aes(colour = gender),
           layout.alg = "fruchte
```



How to plot: Try different layout algorithms

Follow links in ?geom_net for more examples:

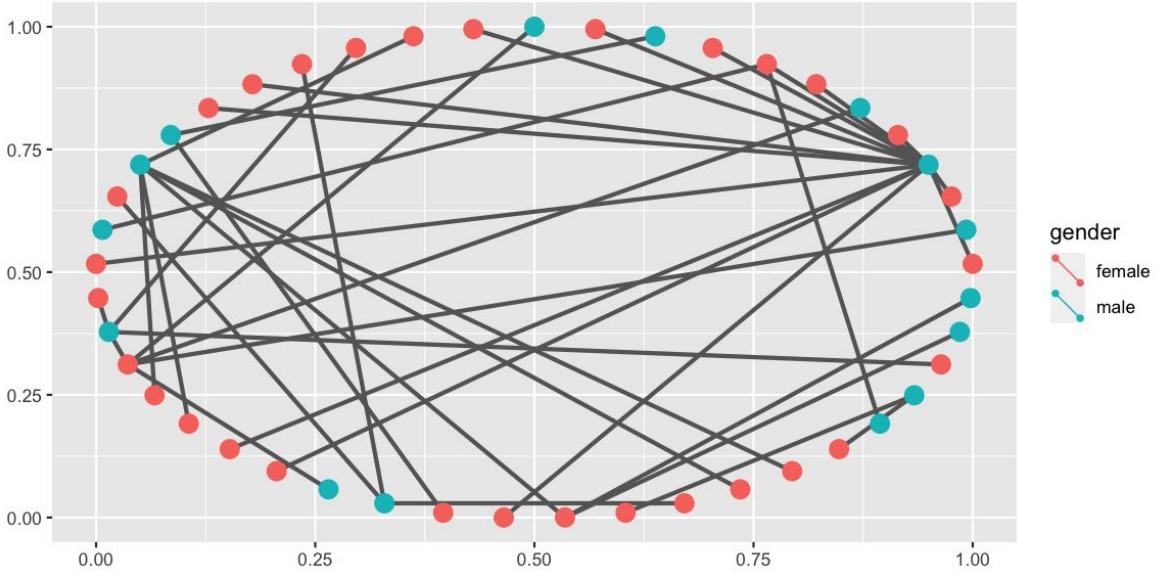
```
set.seed(5556677)
ggplot(data = madmen_net,
       aes(from_id = from_id,
           to_id = to_id)) +
  geom_net(aes(colour = gender),
           layout.alg = "target"
```



How to plot: Try different layout algorithms

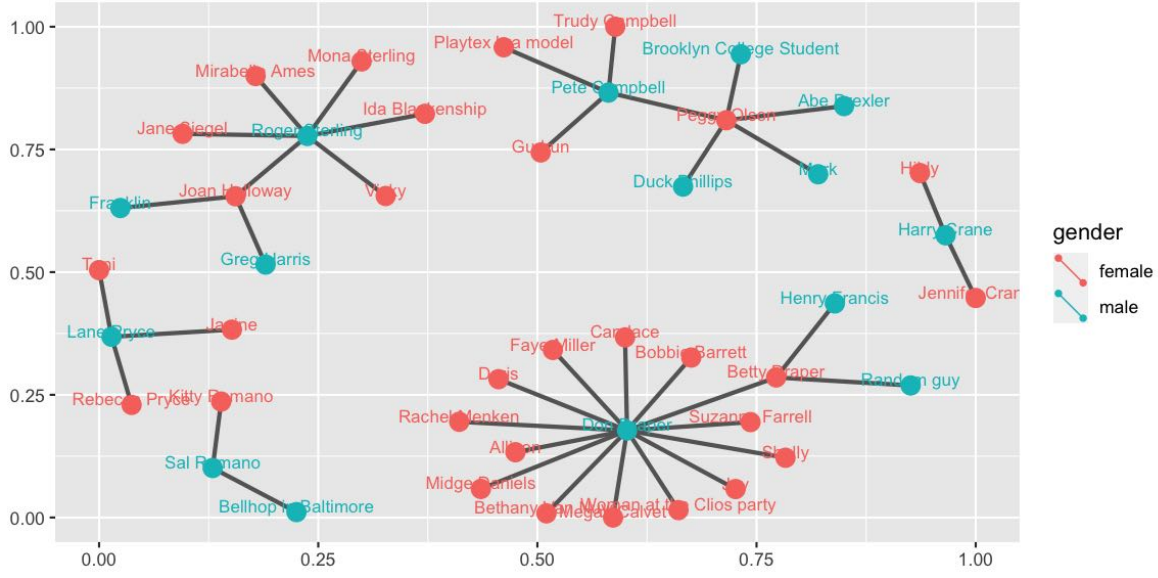
Follow links in ?geom_net for more examples:

```
set.seed(5556677)
ggplot(data = madmen_net,
       aes(from_id = from_id,
           to_id = to_id)) +
  geom_net(aes(colour = gender),
           layout.alg = "circle"
```



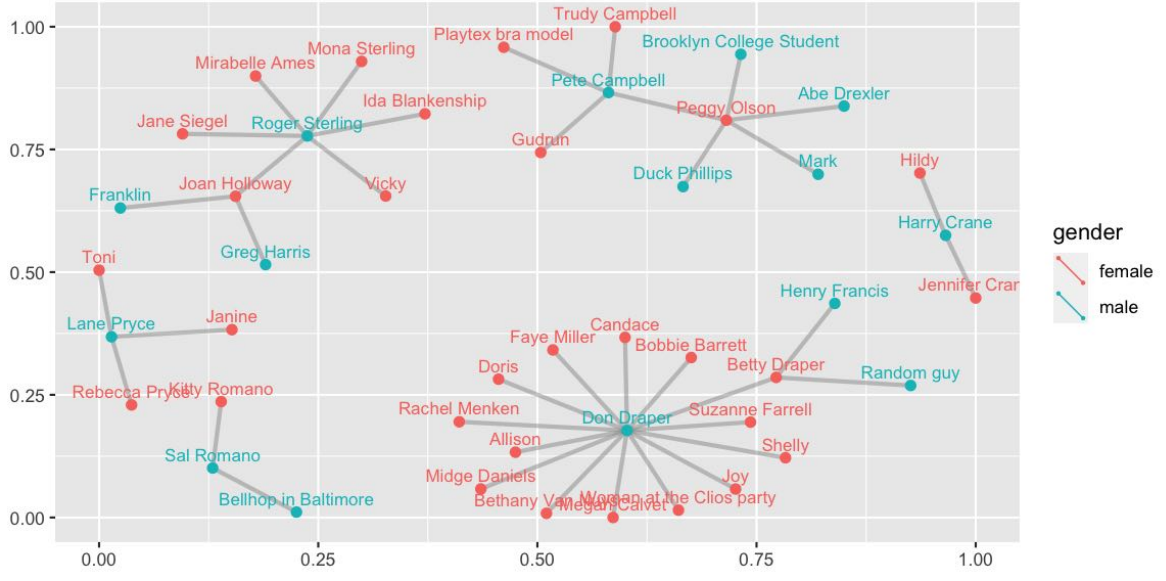
How to plot: Add some labs and decrease font

```
set.seed(5556677)
ggplot(data = madmen_net,
       aes(from_id = from_id,
           to_id = to_id)) +
  geom_net(aes(colour = gender),
           layout.alg = "kamadak
           directed = FALSE,
           labelon = TRUE,
           fontsize = 3)
```



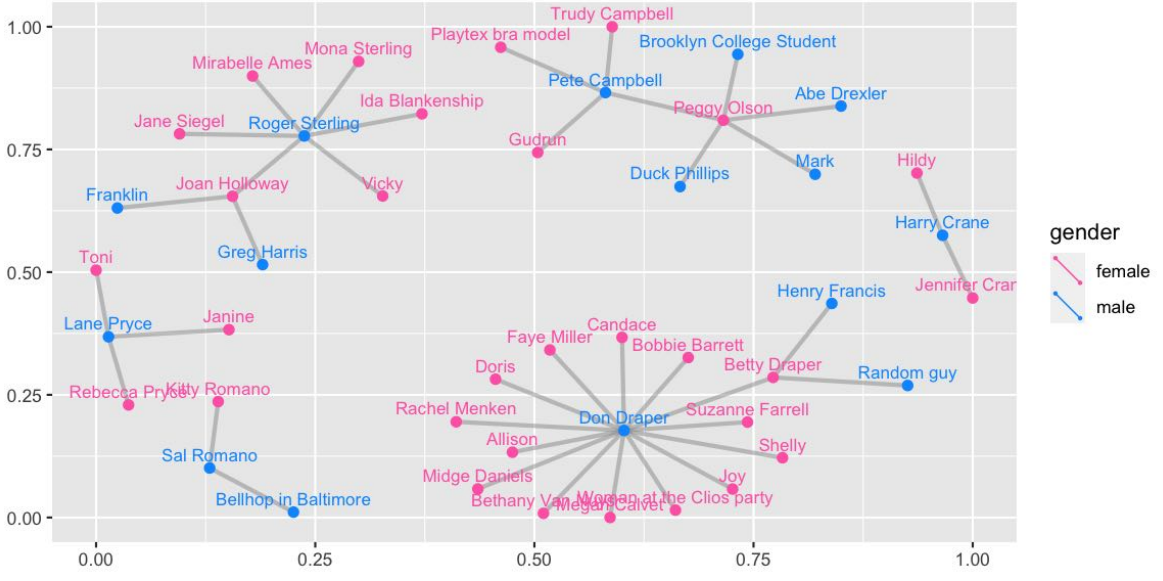
How to plot: Change edge colour/size

```
set.seed(5556677)
ggplot(data = madmen_net,
       aes(from_id = from_id,
           to_id = to_id)) +
  geom_net(aes(colour = gender),
           layout.alg = "kamadak
           directed = FALSE,
           labelon = TRUE,
           fontsize = 3,
           size = 2,
           vjust = -0.6,
           ecolour = "grey60",
           ealpha = 0.5)
```



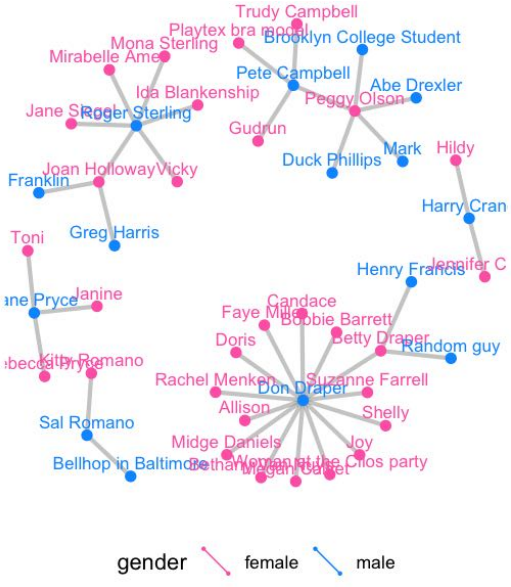
How to plot: Add colours + theme

```
set.seed(5556677)
ggplot(data = madmen_net,
       aes(from_id = from_id,
           to_id = to_id)) +
  geom_net(aes(colour = gender),
           layout.alg = "kamadak
           directed = FALSE,
           labelon = TRUE,
           fontsize = 3,
           size = 2,
           vjust = -0.6,
           ecolour = "grey60",
           ealpha = 0.5) +
    scale_colour_manual(
      values = c("#FF69B4", "#0
      )
```



How to plot: Add theme + move legend

```
set.seed(5556677)
gg_madmen_net <-
ggplot(data = madmen_net,
       aes(from_id = from_id,
           to_id = to_id)) +
  geom_net(aes(colour = gender),
           layout.alg = "kamadak
           directed = FALSE,
           labelon = TRUE,
           fontsize = 3,
           size = 2,
           vjust = -0.6,
           ecolour = "grey60",
           ealpha = 0.5) +
    scale_colour_manual(values =
  theme_net() +
  theme(legend.position = "botto
gg_madmen_net
```



Which character was most connected?

```
madmen_edges
## # A tibble: 39 x 2
##    Name1         Name2
##    <chr>         <chr>
##  1 Betty Draper Henry Francis
##  2 Betty Draper Random guy
##  3 Don Draper   Allison
##  4 Don Draper   Bethany Van Nuys
##  5 Don Draper   Betty Draper
##  6 Don Draper   Bobbie Barrett
##  7 Don Draper   Candace
##  8 Don Draper   Doris
##  9 Don Draper   Faye Miller
## 10 Don Draper   Joy
## # … with 29 more rows
```

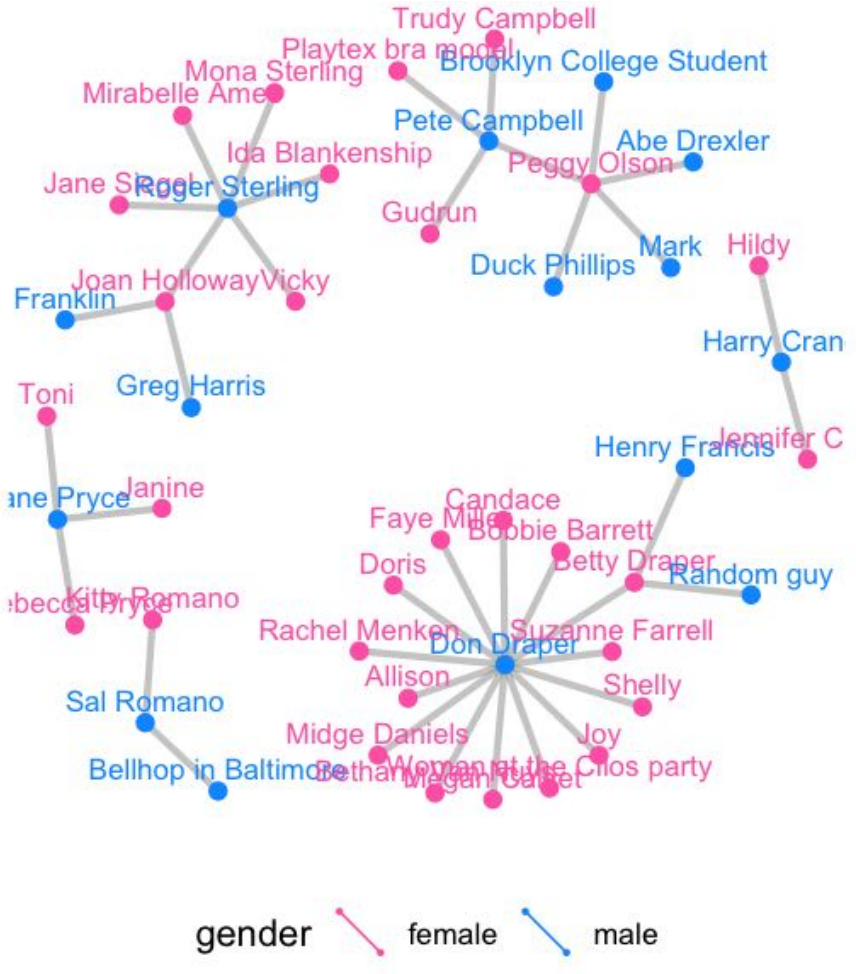
Which character was most connected?

```
madmen_edges %>%
  pivot_longer(cols = c(Name1, Name2),
               names_to = "List",
               values_to = "Name")
## # A tibble: 78 x 2
##    List  Name
##    <chr> <chr>
##  1 Name1 Betty Draper
##  2 Name2 Henry Francis
##  3 Name1 Betty Draper
##  4 Name2 Random guy
##  5 Name1 Don Draper
##  6 Name2 Allison
##  7 Name1 Don Draper
##  8 Name2 Bethany Van Nuys
##  9 Name1 Don Draper
## 10 Name2 Betty Draper
## # … with 68 more rows
```

Which character was most connected?

```
madmen_edges %>%
  pivot_longer(cols = c(Name1, Name2),
               names_to = "List",
               values_to = "Name") %>%
  count(Name, sort = TRUE)
## # A tibble: 45 x 2
##    Name               n
##    <chr>          <int>
##  1 Don Draper        14
##  2 Roger Sterling     6
##  3 Peggy Olson        5
##  4 Pete Campbell      4
##  5 Betty Draper       3
##  6 Joan Holloway      3
##  7 Lane Pryce         3
##  8 Harry Crane        2
##  9 Sal Romano         2
## 10 Abe Drexler        1
## # … with 35 more rows
```

Which character was most connected?



What do we learn?

- Joan Holloway had a lot of affairs, all with loyal partners except for his wife Betty, who had two affairs herself
- Followed by Woman at Clios party

Your Turn:

- Open 9a-madmen.Rmd
- Replicate the plots used in the lecture
- Explore a few different layout algorithms

Example: American college football

Early American football outfits were like Australian AFL today!

Source: wikicommons

Example: American college football

Fall 2000 Season of Division I college football.

- Nodes are the teams, edges are the matches.
- Teams are broken into "conferences" which are the primary competition, but they can play outside this group.

American college football data: Edges

```
football_edges
## # A tibble: 613 x 4
##    from          to                  same.conf intriad
##    <chr>         <chr>                   <dbl> <lgl>
##  1 BrighamYoung FloridaState                0 TRUE
##  2 Iowa         KansasState                 0 TRUE
##  3 BrighamYoung NewMexico                   1 TRUE
##  4 NewMexico    TexasTech                   0 FALSE
##  5 KansasState  TexasTech                   1 TRUE
##  6 Iowa         PennState                   1 TRUE
##  7 PennState    SouthernCalifornia          0 FALSE
##  8 ArizonaState SouthernCalifornia          1 TRUE
##  9 ArizonaState SanDiegoState               0 TRUE
## 10 BrighamYoung SanDiegoState               1 TRUE
## # … with 603 more rows
```

American college football data: Nodes

```
football_nodes
## # A tibble: 115 x 2
##    label              value
##    <chr>              <chr>
##  1 BrighamYoung       Mountain West
##  2 FloridaState       Atlantic Coast
##  3 Iowa               Big Ten
##  4 KansasState        Big Twelve
##  5 NewMexico          Mountain West
##  6 TexasTech          Big Twelve
##  7 PennState          Big Ten
##  8 SouthernCalifornia Pacific Ten
##  9 ArizonaState       Pacific Ten
## 10 SanDiegoState      Mountain West
## # … with 105 more rows
```

American college football: joining the data

```
# data step: merge vertices and edges
ftnet <- full_join(football_edges,
                   football_nodes,
                   by = c("from" = "label")) %>%
  mutate(schools = if_else(value == "Independents", from, ""))

ftnet
## # A tibble: 621 x 6
##    from         to                same.conf intriad value         schools
##    <chr>        <chr>                 <dbl> <lgl>   <chr>         <chr>
##  1 BrighamYoung FloridaState              0 TRUE    Mountain West ""
##  2 Iowa         KansasState               0 TRUE    Big Ten       ""
##  3 BrighamYoung NewMexico                 1 TRUE    Mountain West ""
##  4 NewMexico    TexasTech                 0 FALSE   Mountain West ""
##  5 KansasState  TexasTech                 1 TRUE    Big Twelve    ""
##  6 Iowa         PennState                 1 TRUE    Big Ten       ""
##  7 PennState    SouthernCalifornia        0 FALSE   Big Ten       ""
##  8 ArizonaState SouthernCalifornia        1 TRUE    Pacific Ten   ""
##  9 ArizonaState SanDiegoState             0 TRUE    Pacific Ten   ""
## 10 BrighamYoung SanDiegoState             1 TRUE    Mountain West ""
```

American college football: Identify ndoes

```
ggplot(data = ftnet,
       aes(from_id = from, to_id = to)) +
  geom_net(
    aes(colour = value,
        group = value,
        linetype = factor(1-same.conf),
        label = schools),
    linewidth = 0.5,
    size = 5,
    vjust = -0.75,
    alpha = 0.3,
    layout.alg = 'fruchtermanreingold'
  ) +
  theme_net() +
  theme(legend.position = "bottom") +
  scale_colour_brewer("Conference", palette = "Paired")
```

American college football: Add colours and linetypes

```
ggplot(data = ftnet,
       aes(from_id = from, to_id = to)) +
  geom_net(
    aes(colour = value,
        group = value,
        linetype = factor(1-same.conf),
        label = schools),
    linewidth = 0.5,
    size = 5,
    vjust = -0.75,
    alpha = 0.3,
    layout.alg = 'fruchtermanreingold'
  ) +
  theme_net() +
  theme(legend.position = "bottom") +
  scale_colour_brewer("Conference", palette = "Paired")
```

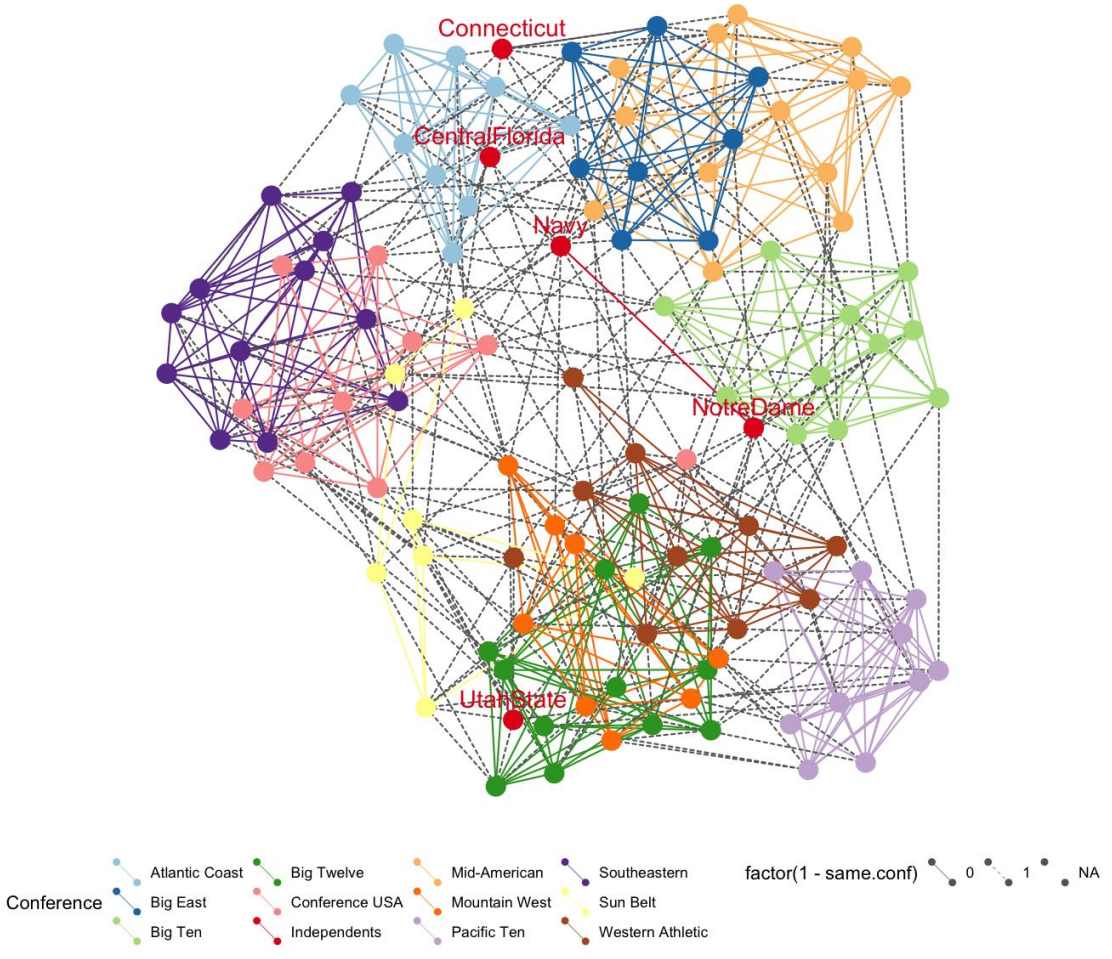
American college football: Line features

```
ggplot(data = ftnet,
       aes(from_id = from, to_id = to)) +
  geom_net(
    aes(colour = value,
        group = value,
        linetype = factor(1-same.conf),
        label = schools),
    linewidth = 0.5,
    size = 5,
    vjust = -0.75,
    alpha = 0.3,
    layout.alg = 'fruchtermanreingold'
  ) +
  theme_net() +
  theme(legend.position = "bottom") +
  scale_colour_brewer("Conference", palette = "Paired")
```

American college football: Theme features and colours

```
ggplot(data = ftnet,
       aes(from_id = from, to_id = to)) +
  geom_net(
    aes(colour = value,
        group = value,
        linetype = factor(1-same.conf),
        label = schools),
    linewidth = 0.5,
    size = 5,
    vjust = -0.75,
    alpha = 0.3,
    layout.alg = 'fruchtermanreingold'
  ) +
  theme_net() +
  theme(legend.position = "bottom") +
  scale_colour_brewer("Conference", palette = "Paired")
```

American college football:



What do we learn?

- Remember layout is done to place nodes that are more similar close together in the display.
- The colours indicate conference the team belongs too. For the most part, conferences are clustered, more similar to each other than other conferences.
- There are some clusters of conference groups, eg Mid-American, Big East, and Atlantic Coast
- The Independents are independent
- Some teams play far afield from their conference.

Our Turn: Harry Potter characters

See "9a-harry-potter.Rmd"

Source: *wikicommons*

Example: Harry Potter characters

There is a connection between two students if one provides emotional support to the other at some point in the book.

- Code to pull the data together is provided by Sam Tyner here.

Harry potter data as nodes and edges

```
hp_all
## # A tibble: 720 x 6
##    book  from_id           to_id               schoolyear gender house
##    <chr> <chr>             <chr>                    <dbl> <chr>  <chr>
##  1 1     Dean Thomas       Harry James Potter        1991 M      Gryffindor
##  2 1     Dean Thomas       Hermione Granger          1991 M      Gryffindor
##  3 1     Dean Thomas       Neville Longbottom        1991 M      Gryffindor
##  4 1     Dean Thomas       Ronald Weasley            1991 M      Gryffindor
##  5 1     Dean Thomas       Seamus Finnigan           1991 M      Gryffindor
##  6 1     Fred Weasley      George Weasley            1989 M      Gryffindor
##  7 1     Fred Weasley      Harry James Potter        1989 M      Gryffindor
##  8 1     George Weasley    Fred Weasley              1989 M      Gryffindor
##  9 1     George Weasley    Harry James Potter        1989 M      Gryffindor
## 10 1     Harry James Potter Dean Thomas              1991 M      Gryffindor
## # … with 710 more rows
```

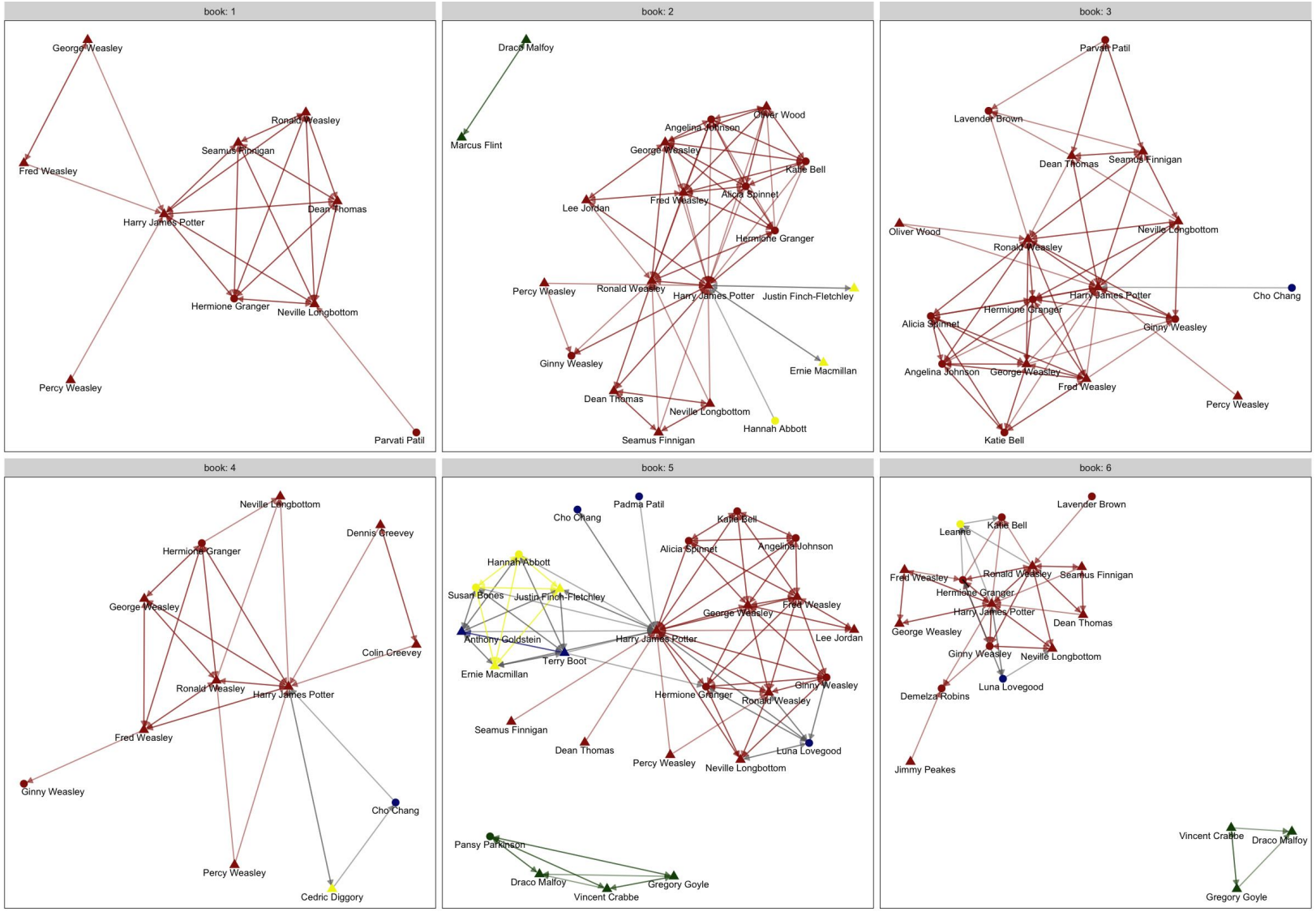
Let's plot the characters

```
ggplot(data = hp_all,
       aes(from_id = from_id,
           to_id = to_id)) +
  geom_net(aes(colour = house, group = house, shape = gender),
           fiteach=T,
           directed = T,
           size = 3,
           linewidth = .5,
           ealpha = .5,
           labelon = T,
           fontsize = 3,
           repel = T,
           labelcolour = "black",
           arrowsize = .5,
           singletons = FALSE) +
  scale_colour_manual(values = c("#941B08","#F1F31C", "#071A80", "#154C07")) +
  facet_wrap(~book, labeller = "label_both", ncol=3) +
  theme_net() +
  theme(panel.background = element_rect(colour = 'black'),
        legend.position="bottom")
```

Some more questions

- In the first book, which characters had the most connections?
- How about the least connections?

Let's plot the characters



Summary

- To make a network analysis, you need:
- an association matrix, that describes how nodes (vertices) are connected to each other
- a layout algorithm to place the nodes optimally so that the fewest edges cross, or that the nodes that are most closely associated are near to each other.

Your turn: rstudio exercise

- Complete 9a-class.Rmd
- Read in last semesters class data, which contains

  `s1_name` and `s2_name` are the first names of class members, and tutors, with the latter being the "go-to" person for the former.
- Write the code to produce a class network that looks something like the plot on the right.