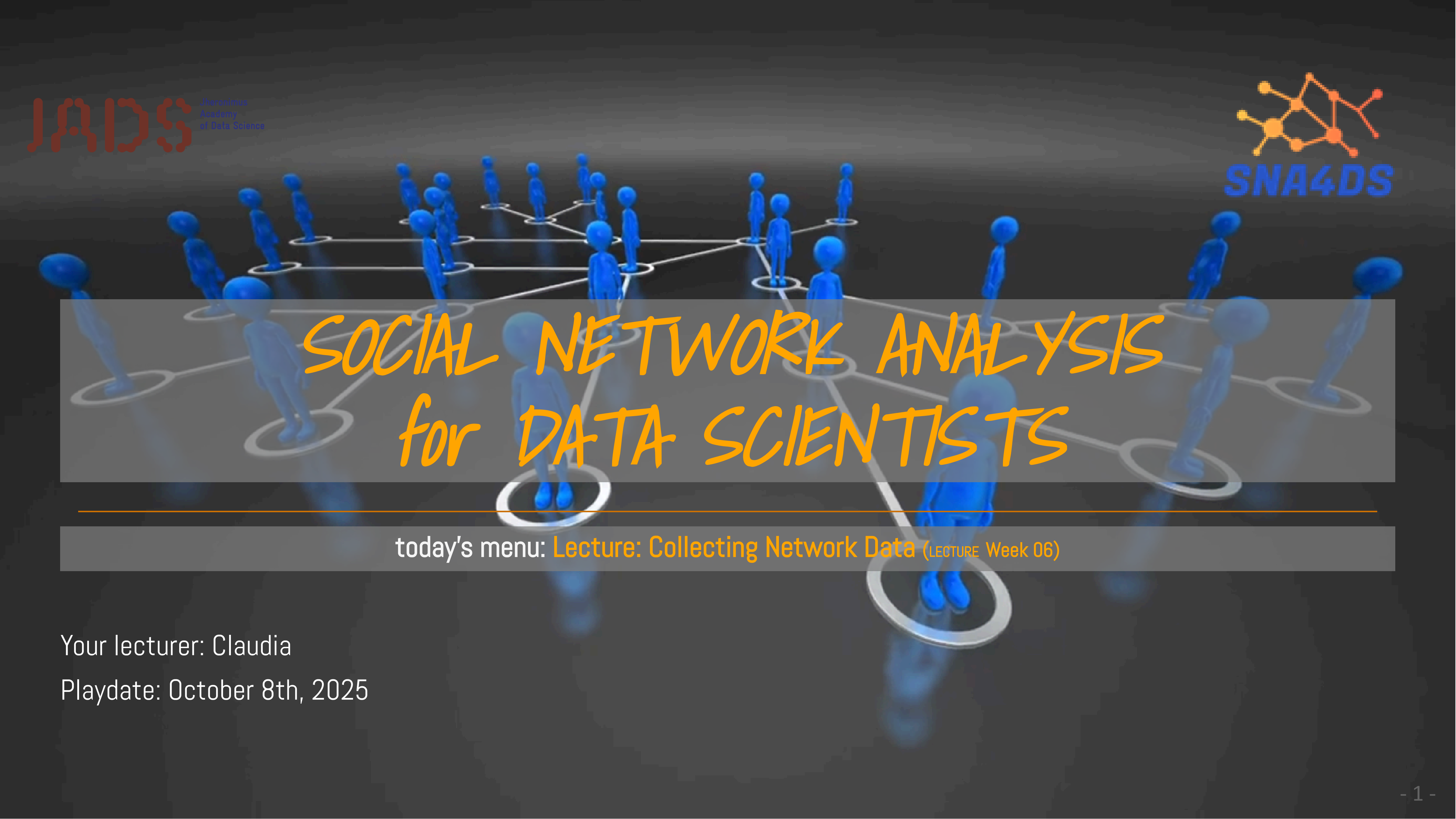


SOCIAL NETWORK ANALYSIS for DATA SCIENTISTS

today's menu: Lecture: Collecting Network Data (LECTURE Week 06)

Your lecturer: Claudia

Playdate: October 8th, 2025

Let's collect some data!

canvas -> modules -> week 6 -> link

https://tilburghumanities.eu.qualtrics.com/jfe/form/SV_e9YKCcujyVw8iPA

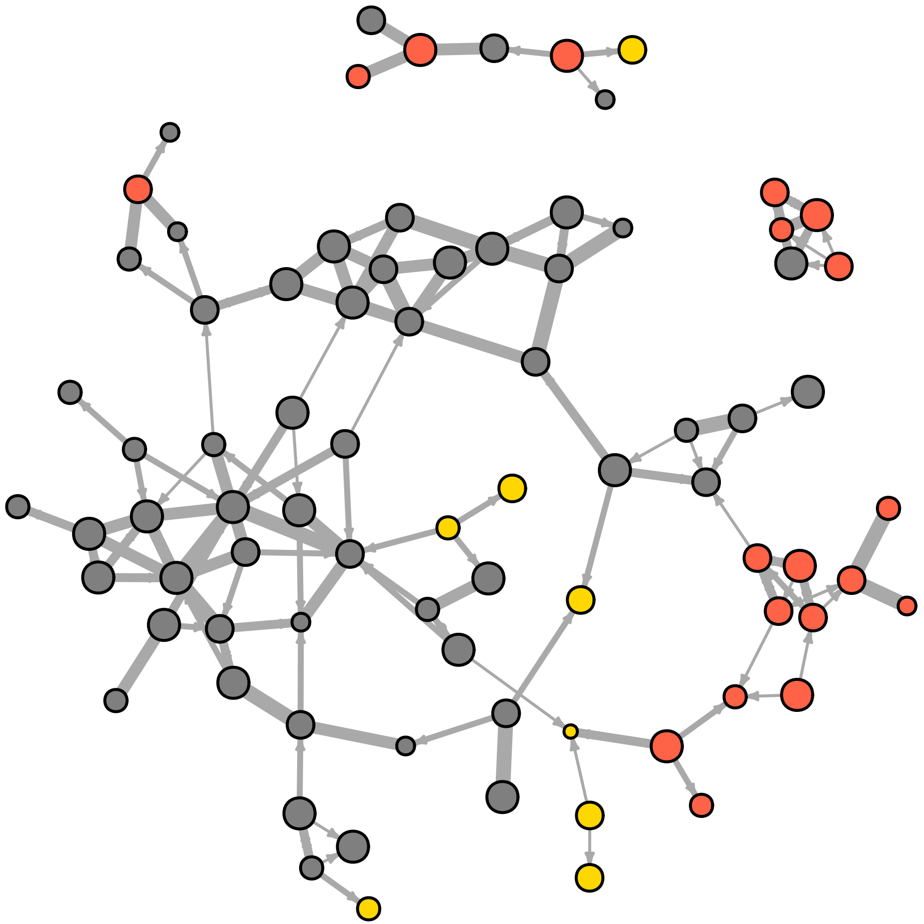
Agenda for today's lecture

1) Think and Reflect on the Survey Questions

2) Explore network data collection strategies

3) Discuss how do you process newly collected data

This are the relationships in the cohort of three years ago



1) Think and Reflect on the Survey Questions

Think and Reflect

- Q1 Why are these questions suitable for collecting network data?
- Q2 Which question collects what kind of data?
- Q3 What could be a bias for this data collection?
- Q4 What are the possible ethical concerns that might arise?

Overview of the Survey questions

This survey is about understanding how to collect data

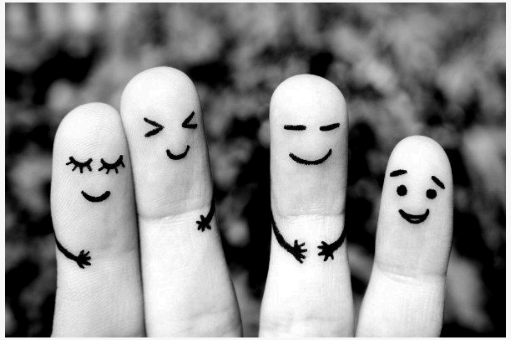
- Can you please select your name from the list below?

- How much do you like data science?

- Would you please select the name of the one coursemate with whom you interact the most?

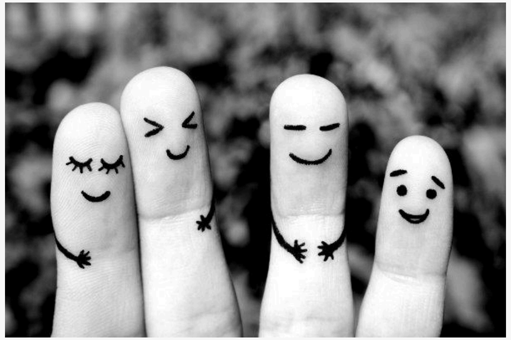
- How often do you interact?

Q1 Why are these questions suitable for collecting network data?



Q1 Why are these questions suitable for collecting network data?

This questions are suitable for collecting network data about friendship
not every kind of network data



- collecting info about friendship (RELATIONSHIP - network)
- collecting information about respondents (RESPONDENTS ATTRIBUTES)
- collecting information about the relationship between pair of respondents (RELATIONSHIP ATTRIBUTE)

There is no universally valid questionnaire

Q2 Which question collects what data?

Q2 Which question collects what data?

Levels

- network info
- node attribute
- edge attribute

Mentimeter!

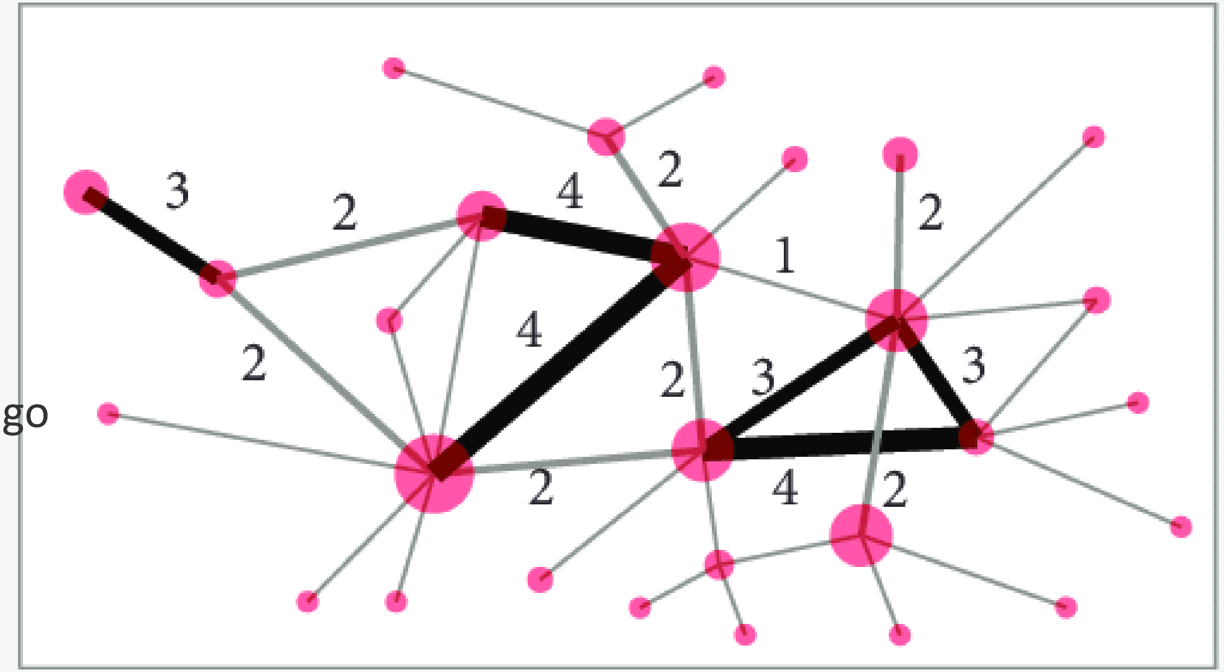
7268 6485

www.menti.com

https://www.menti.com/al3jcf1aiwgp

"Types" of data

- Network Attribute
  - respondent/ego
  - alter
- node Attribute: attributes of the respondent/ego
- edge Attribute: attributes of the relationship



You collect what you need to answer your research question.

Examples of research questions

Which types of data you need?

Examples of research questions

Which types of data you need?

Does the gender affect the likelihood of pupils playing together in kindergarten?

Examples of research questions

Which types of data you need?

Does the gender affect the likelihood of pupils playing together in kindergarten?

Network

Node attribute

Examples of research questions

Which types of data you need?

Does the gender affect the likelihood of pupils playing together in kindergarten?

Network

Node attribute

Are people that retweet each other more likely to have the same opinion?

Examples of research questions

Which types of data you need?

Does the gender affect the likelihood of pupils playing together in kindergarten?

Network

Node attribute

Are people that retweet each other more likely to have the same opinion?

Network (retweet)

Edge Attribute (opinion in the tweet - categorical)

Q3 What could be a bias for this data collection?

I left a few "problems" in the survey, let's see if you catch them

Q3 What could be a bias for this data collection?

I left a few "problems" in the survey, let's see if you catch them

The number of friends.

Is 3 a good number?

Q3 What could be a bias for this data collection?

I left a few "problems" in the survey, let's see if you catch them

The number of friends.

Is 3 a good number?

What is a friend?

Are these questions capturing 'friendship'? Is a colleague a friend?

"To what extent do you consider this person your friend compared to your friends outside uni?" it could mean too many things

Maybe this is fine, but we need to make sure to define what we are measuring before

unbiased design

Think carefully

- Define your target relationship

- Decide what is the right number of connections per respondent to collect

- Make sure your research design is linked to the research question

- Ask the questions in a specific way

  - make sure you do not ask two questions in one.

- Make sure that you end up with measurable answers

  - preferably use Likert scales with odd numbers of points (3, 5, 7, 11)

Remember: A fully unbiased study does not exist!

Sampling

- Do you have access to the entire population?

yes? Run a survey like the one we used

- targeted a group with a boundary (SNA4DS students)
- Not everyone responded, but this is simply missing data

No? You need to use an egonet approach

You ask them to provide the name of their relationship since you might not even know how many are they

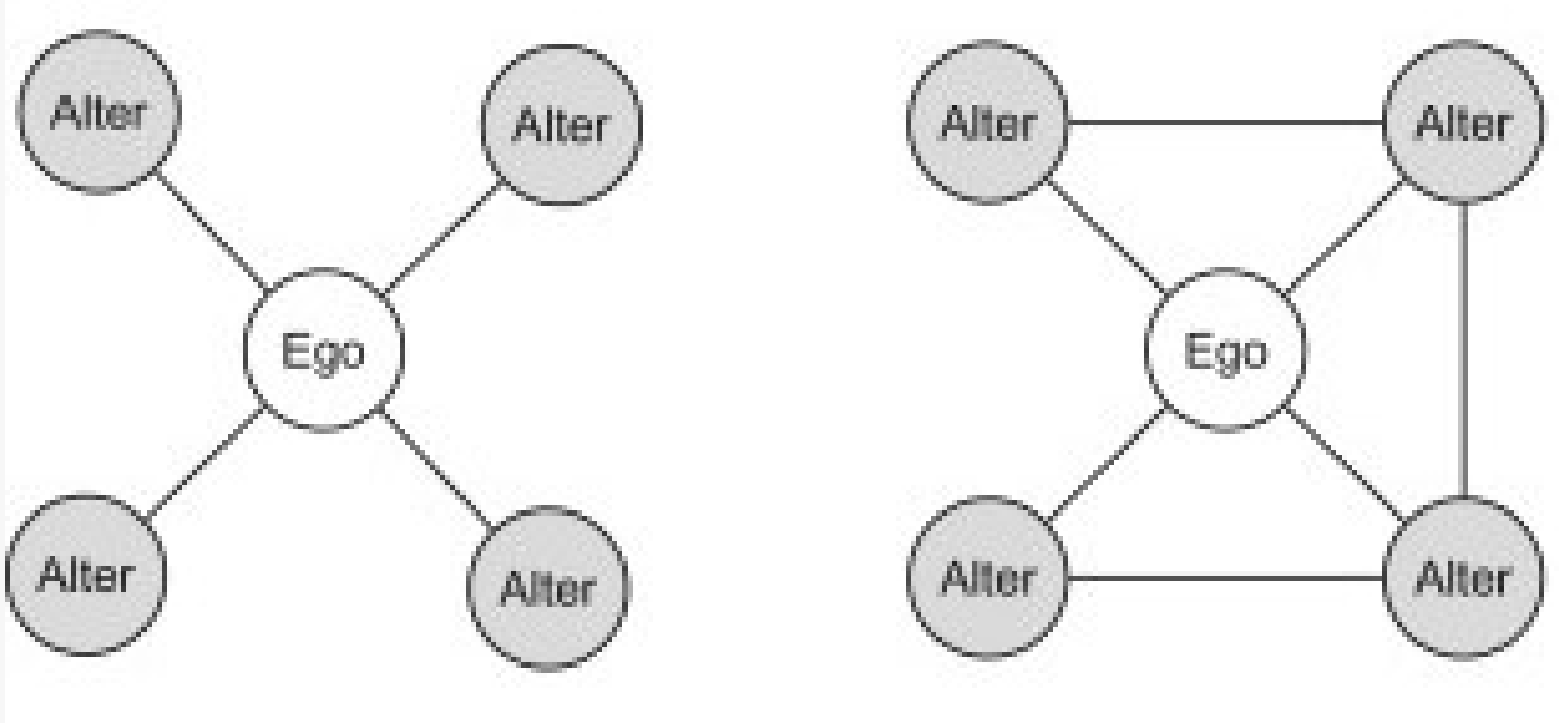
entire sample: Roster with names

egonet: "snowballing" data

Access to the entire population with missing data DIFFERS from access to a sample

Egonets simple

- Who are your friends?
- Are they friends among each other?



Egonets advanced design

- You can tailor the design to your needs

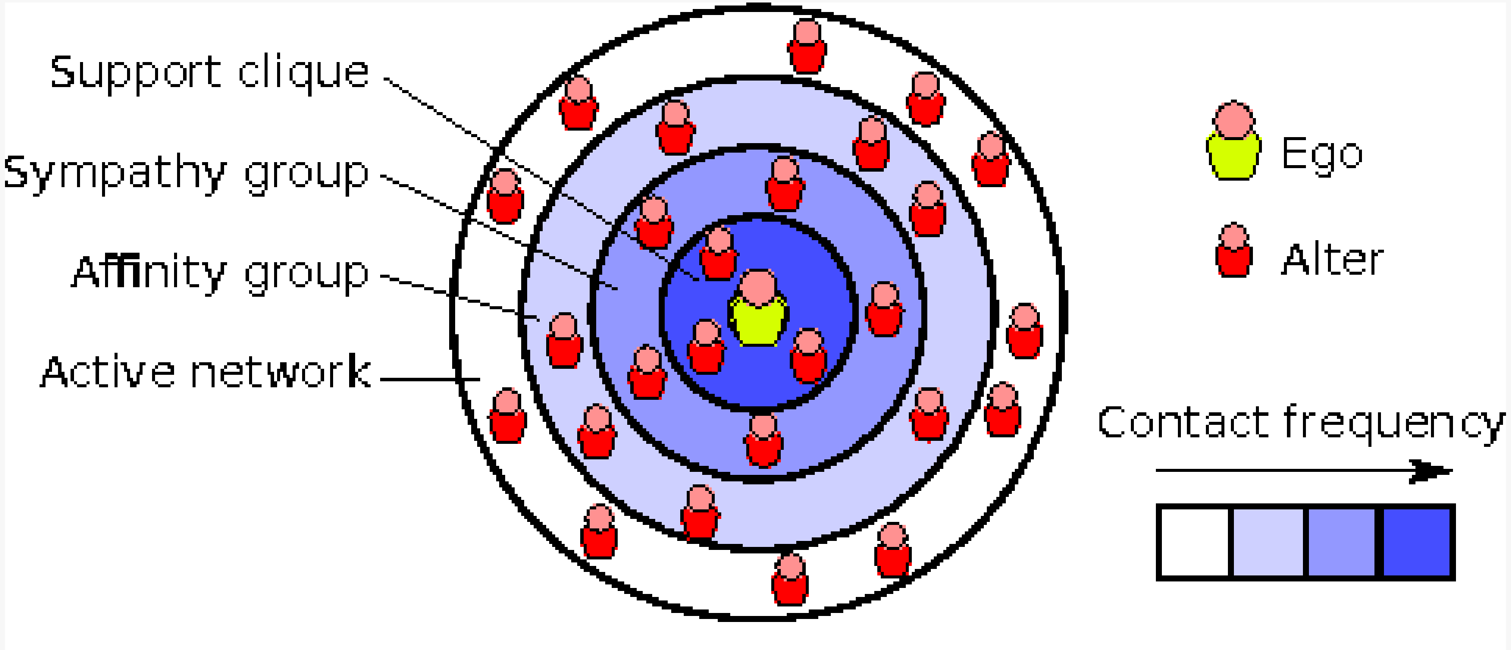


Fig. 1. Ego network model.

Q4 What are the possible ethical concerns that might arise?

Q4 What are the possible ethical concerns that might arise?

- Anonyminity and confidentiality
  - network data is particularly sensitive on this perspective
  - you need to anonymize it and store it properly

Q4 What are the possible ethical concerns that might arise?

- Anonyminity and confidentiality
  - network data is particularly sensitive on this perspective
  - you need to anonymize it and store it properly
- Sensitive questions
  - such as gender, or medical information

Q4 What are the possible ethical concerns that might arise?

- Anonyminity and confidentiality
  - network data is particularly sensitive on this perspective
  - you need to anonymize it and store it properly
- Sensitive questions
  - such as gender, or medical information
- Be aware of your intended sample and protect it
  - Are you recruiting a sensitive sample? Make sure no one is in danger (e.g. AIDS patients)
  - make sure no one is recognizable even 10 years later

Q4 What are the possible ethical concerns that might arise?

- Anonyminity and confidentiality
  - network data is particularly sensitive on this perspective
  - you need to anonymize it and store it properly
- Sensitive questions
  - such as gender, or medical information
- Know who is your intended sample and protect it
  - Are you recruiting a sensitive sample? Make sure no one is in danger (e.g. AIDS patients)
  - make sure no one is recognizable even 10 years later
- GDPR
  - Get consent and agreement

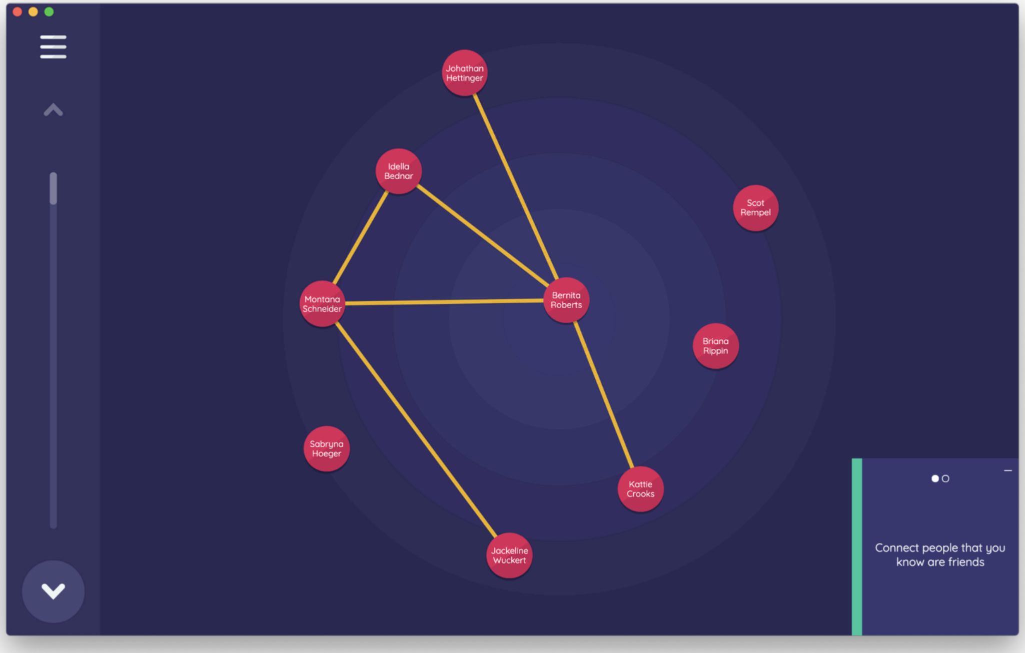
2) Explore network data collection strategies

Sofware

For running a survey you can use regular platforms not specifically meant for networks (Qualtrics, Survey Monkey etc...)

But there are software specifically designed to collect network data. For instance Network Canvas

ego or entire sample



Connect people that you know are friends

M. Birkett, J. Melville, P. Janulis, G. Phillips, N. Contractor, B. Hogan, Network Canvas: Key decisions in the design of an interviewer-assisted network data collection software suite, Social Networks, Volume 66, 2021, Pages 114-124.

Data collection strategies for people: online survey + fieldwork

We explored the online Survey data collection

A survey does not get you everywhere

kids? elderly people?

You might need to do fieldwork

yes paper and pen! or Network canvas

Data collection strategies for people: online data

Spotify: Scrape data.

```
# selected features
str(fsel)
```

```
## 'data.frame':    2 obs. of  39 variables:
##  $ artist_name                 : chr  "100 gecs" "100 gecs"
##  $ artist_id                   : chr  "6PfSUFtkMVoDkx4MQkzOi3" "6PfSUFtkMVoDkx4MQkzOi3"
##  $ album_id                    : chr  "0qnExDZfz0kVeBjixPsyjS" "0qnExDZfz0kVeBjixPsyjS"
##  $ album_type                  : chr  "album" "album"
##  $ album_images                :List of 2
##   ..$ :'data.frame':    3 obs. of  3 variables:
##   .. ..$ height: int  640 300 64
##   .. ..$ url   : chr  "https://i.scdn.co/image/ab67616d0000b273f91a3040f0be854026ad2dd0" "https://i.scd
##   .. ..$ width : int  640 300 64
##   ..$ :'data.frame':    3 obs. of  3 variables:
##   .. ..$ height: int  640 300 64
##   .. ..$ url   : chr  "https://i.scdn.co/image/ab67616d0000b273f91a3040f0be854026ad2dd0" "https://i.scd
##   .. ..$ width : int  640 300 64
##  $ album_release_date          : chr  "2020-07-10" "2020-07-10"
##  $ album_release_year          : num  2020 2020
```

Spotify: Where is the network?

```
# artist one

# ego
fsel$artist_name[1]
```

```
## [1] "100 gecs"
```

```
# alter (list nested into data frame)
str(fsel$artists[1])
```

```
## List of 1
##  $ :'data.frame':    2 obs. of  6 variables:
##   ..$ href                  : chr [1:2] "https://api.spotify.com/v1/artists/6PfSUFtkMVoDkx4MQkzOi3" "htt
##   ..$ id                    : chr [1:2] "6PfSUFtkMVoDkx4MQkzOi3" "335TWGWGFan4vaacJzSiU8"
##   ..$ name                  : chr [1:2] "100 gecs" "A. G. Cook"
##   ..$ type                  : chr [1:2] "artist" "artist"
##   ..$ uri                   : chr [1:2] "spotify:artist:6PfSUFtkMVoDkx4MQkzOi3" "spotify:artist:335TWGWG
```

You loop through to extract the information that you need

Other examples of online sources of network data: Twitter (old data), Instagram, Facebook, ...

Data collection strategies for alive creatures

Wearable sensors VS direct observation

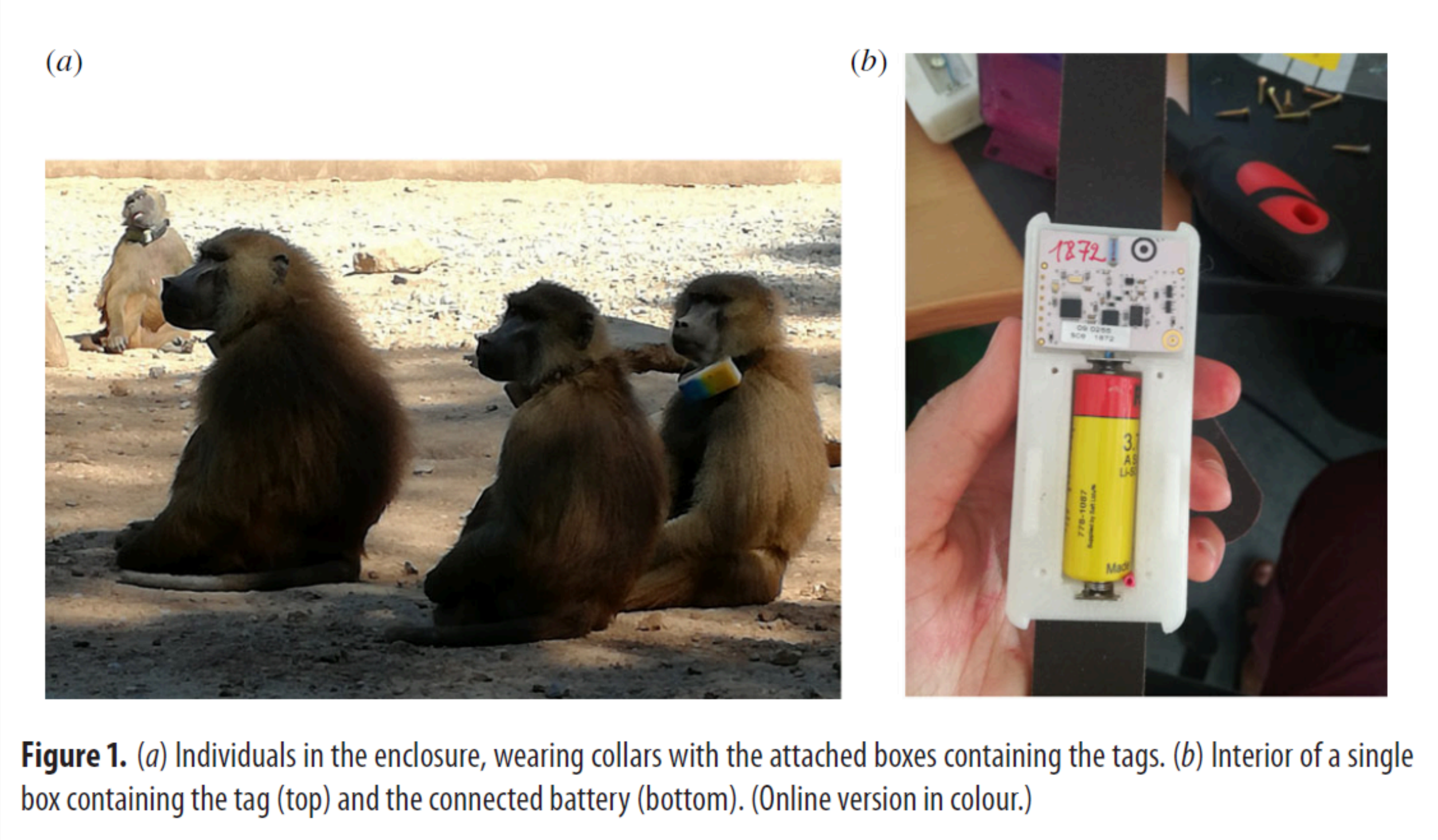


Figure 1. (*a*) Individuals in the enclosure, wearing collars with the attached boxes containing the tags. (*b*) Interior of a single box containing the tag (top) and the connected battery (bottom). (Online version in colour.)

Gelardi V, Godard J, Paleressompoulle D, Claidiere N, Barrat A. 2020 Measuring social networks in primates: wearable sensors versus direct observations. Proc. R. Soc. A 476: 20190737.
http://dx.doi.org/10.1098/rspa.2019.0737

Data collection strategies for not alive creatures

- Private info: Questionnaire (Service Networks)
- Public info: Observation (Dutch train line or internet network)

Figure 9.2: *Joint drop-ins or group work*

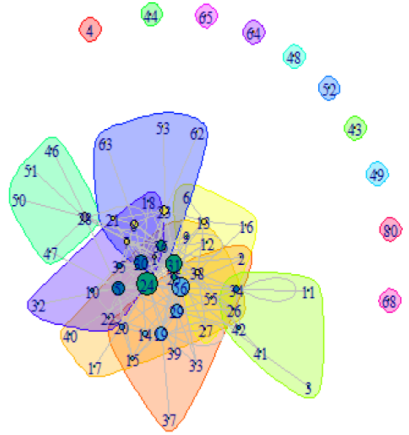
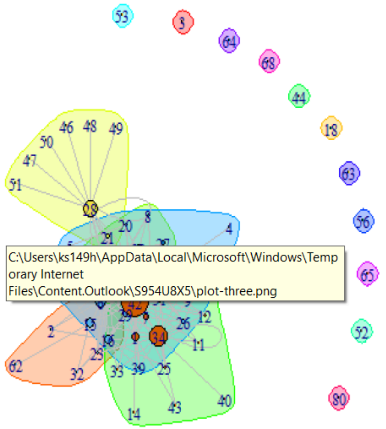


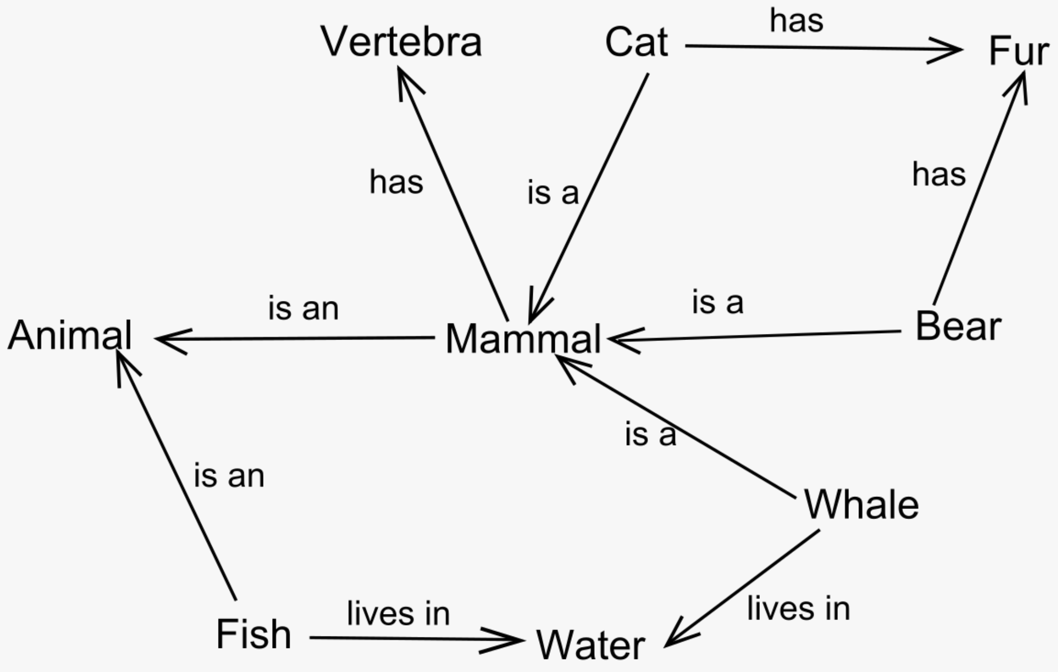
Figure 9.3: *Joint appointments, home visits, and outreach*


C:\Users\ks149h\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.Outlook\S954U8X5\plot-three.png

Clarke, Z. et al., 2019, Needs Assessment: Substance Use, Blood Borne Virus & Sexual Health in Midlothian & East Lothian, NHS Lothians & MELDAP

Data collection strategies for computational networks

Semantic networks (it does not "exist", it is a "theoretical" concept)



Summary of data collection strategies

If possible:

Observe ---> public available data more or less hard to access

[in order of complexity]

- just observe something or someone
- scrape
- use text analysis
- use sensors
- ...

If NOT possible to observe:

Ask questions ---> private information

[in order of complexity]

- Online
- fieldwork (MOST COMPLEX AND EXPENSIVE STRATEGY OF THEM ALL)

Data? Where to find it -

- Kaggle

- Harvard DataVerse

- SNAP Stanford

- Pajek Dataset

- UciNet Datasets

3) Discuss how do you process newly collected data

How do you process newly collected data?

Your data looks like this (you want your data to look like this):

```
survey
```

```
##   respondent alter1 alter2 gender cinemaW1 cinemaW2
## 1       Anna  David   Luke      F        1        0
## 2        Ben     NA     NA      M       NA       NA
## 3       Cleo   Anna  David      O        2        2
```

To analyse it as a network we need to process it

Let's explore one way to do it

Creating an edge list

Let's do the process manually first

- match: respondent -- alter 1
- match: respondent -- alter 2
- append them

```
(EdgeList <- data.frame(rbind(cbind(survey$respondent, survey$alter1),
                              cbind(survey$respondent, survey$alter2))))
```

```
##      X1    X2
## 1 Anna David
## 2  Ben    NA
## 3 Cleo  Anna
## 4 Anna  Luke
## 5  Ben    NA
## 6 Cleo David
```

If you have many columns this could get nasty. Better to loop

Edge List via for loop

```r
col <- survey[ , 1:3] #<- select the columns respondent, alter1, alter2

EdgeList <- data.frame()
temp <- data.frame()

for (i in 2:ncol(col)) {
  temp <- cbind(col[, 1], col[, i])
  EdgeList <- rbind(EdgeList, temp)
}

EdgeList
```

```
##      V1    V2
## 1 Anna David
## 2  Ben    NA
## 3 Cleo  Anna
## 4 Anna  Luke
## 5  Ben    NA
## 6 Cleo David
```

Are we done? We are definitely not. An edge list does not consider isolates.

Make a node list

Let's do the process manually: We need to make sure that all the nodes are included, but one time only

```
(NodeList <- unique(c(survey$respondent, survey$alter1, survey$alter2)))
```

```
## [1] "Anna"   "Ben"    "Cleo"   "David"  "NA"     "Luke"
```

```
NodeList <- na.omit(NodeList) # always remove NAs
```

Unless we use also a node list, Ben would disappear from the sample

A for loop would make your life better in this case too

So, now what?

Now we turn it into a network

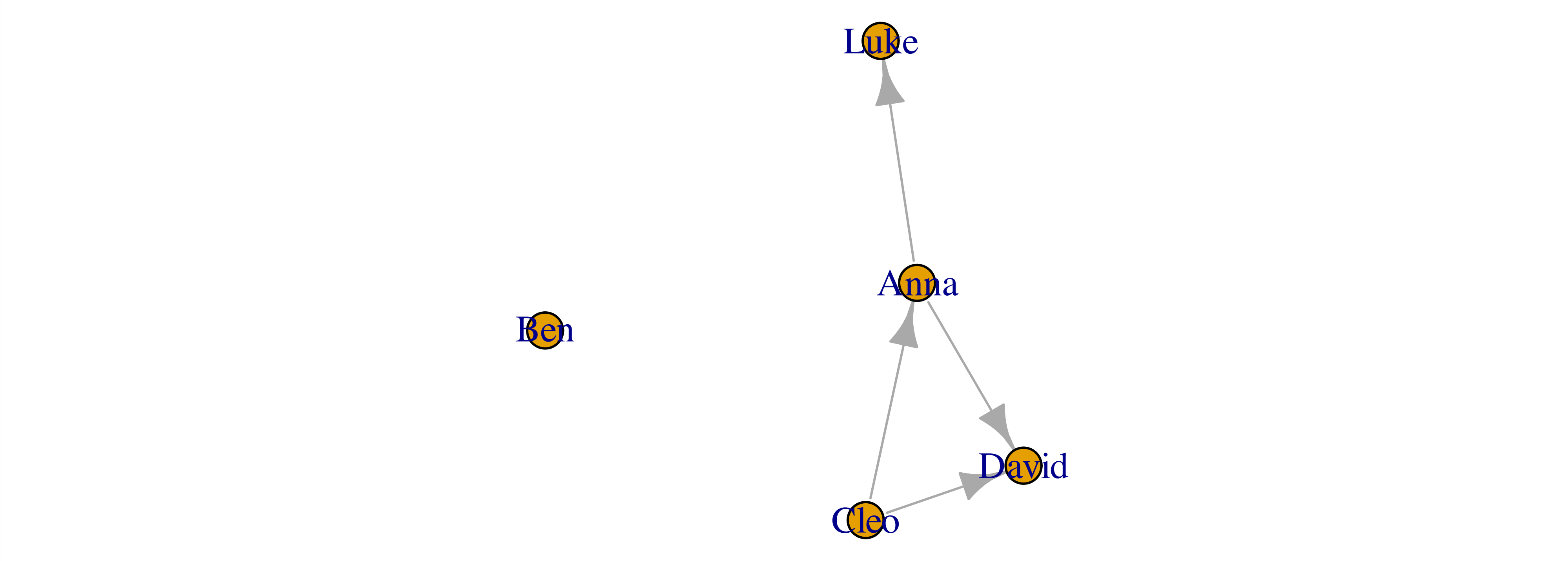
```r
EdgeList <- na.omit(EdgeList)
surveynet <- igraph::graph_from_data_frame(EdgeList, NodeList, directed = TRUE)
```

Disclaimer: there are many other ways, but this is a very useful one

Still, explore the functions in `igraph` and `network` for creating networks

Result

```
plot(surveynet)
```



And the attributes? Edge attributes are processed the same way as the edge list, while node attributes needs to be processed as vectors (consider that you will have missing data for the alters)

To spead up the process

```r
# select columns with network info as df
net <- data.frame(survey[,1:3])

# select columns with edge attributes as df
eAttr <- data.frame(cbind(survey$cinemaW1, survey$cinemaW2))

# prepare a vector with node attribute
nAttr <- survey$gender

# Make the edge list
EdgeList <- snafun::make_edgelist(net, eAttr)
EdgeList <- na.omit(EdgeList)

# Make the node list
NodeList <- snafun::make_nodelist(net, nAttr)

# turn it into a network
surveynet <- igraph::graph_from_data_frame(EdgeList, NodeList, directed = TRUE)
surveynet
```

Enjoy your data hunt!

That's all folks!