# The Young People Survey

# Description of the data set

- Survey
- Carried out in 2013
- Respondents aged 15-30
- Data on general preferences and habits

# Characteristics of the set

- Data format: .csv
- 1010 samples of survey responses
- One data set
- Data from: Kaggle, Young People Survey

# Review of the chosen data mining goals

1. Determining the differences between people with rural vs urban backgrounds
2. Finding the general profile of a ‘money saver’
3. Exploring gender differences in, among others, phobias and fears.

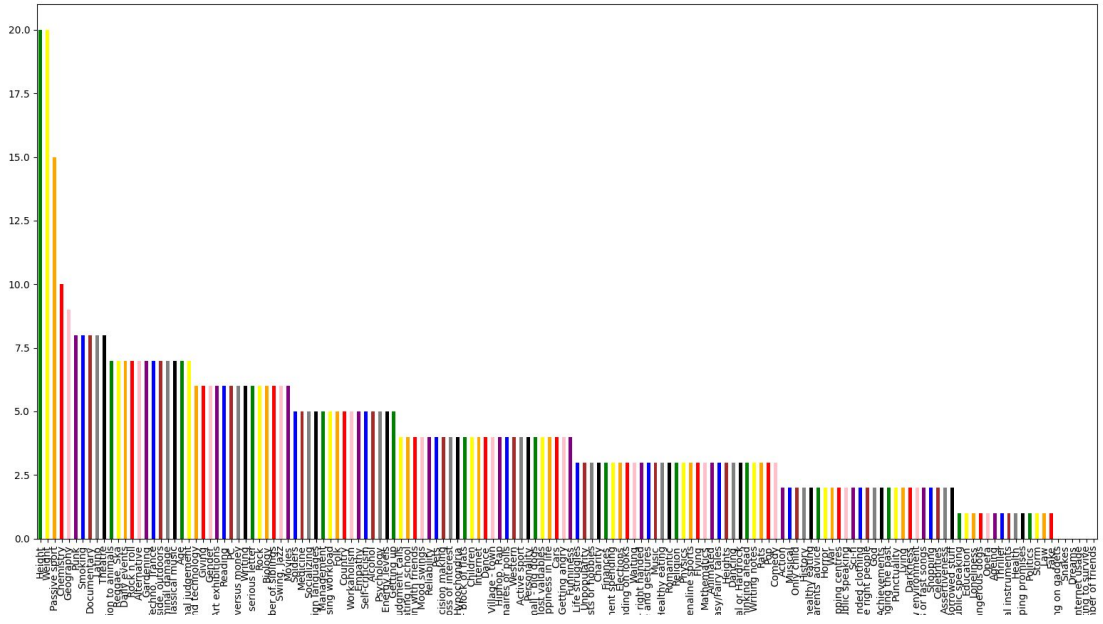
# Discussion of the further steps

For this particular data set we decided that the description model will fit perfectly due to high number and quantitative responses. We decided on clustering algorithm.

Our dataset has many observations and a lot of variable number, that is why we chose Ward’s method.

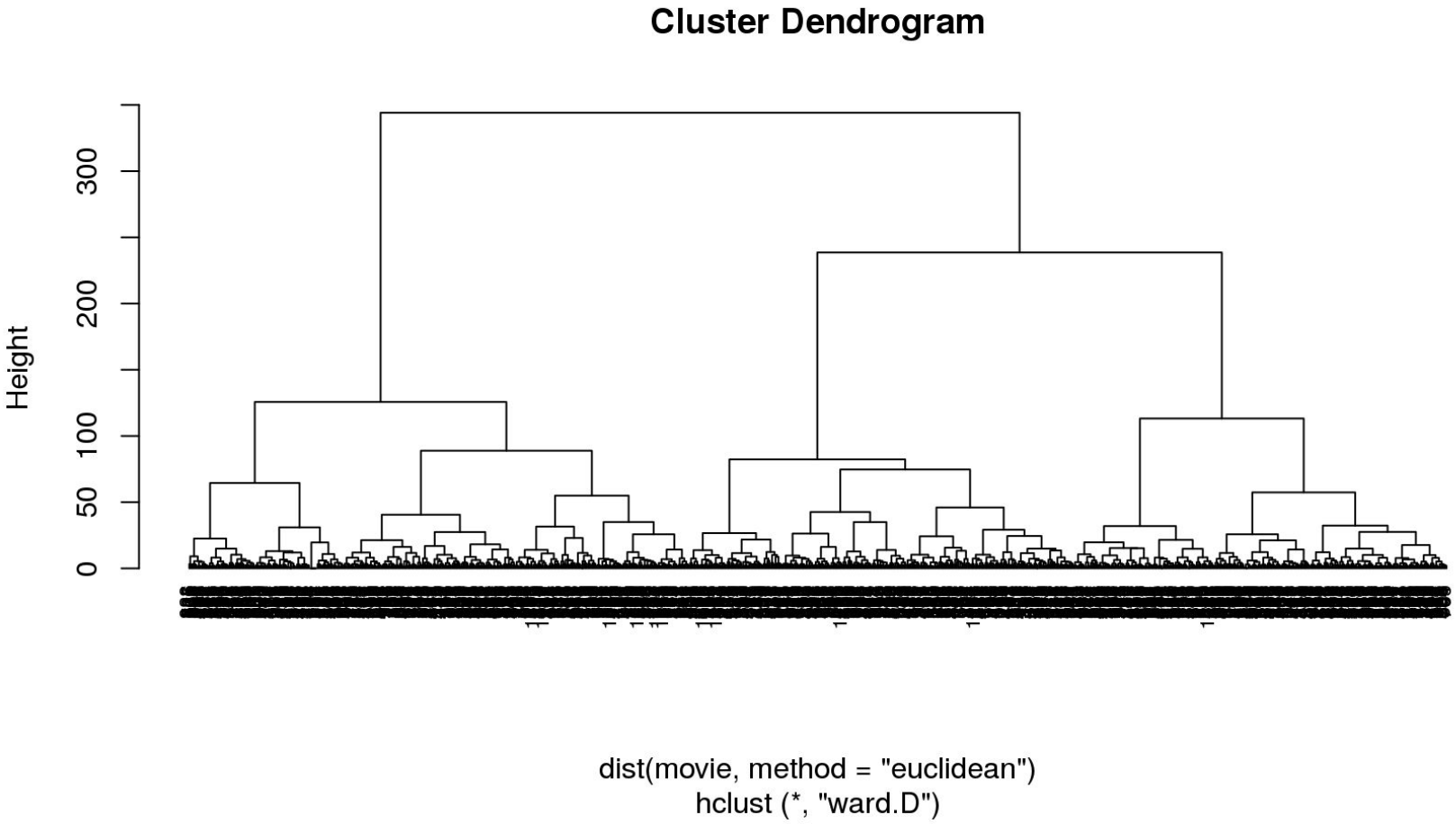
# Description of data preparation

- Missing data handling,

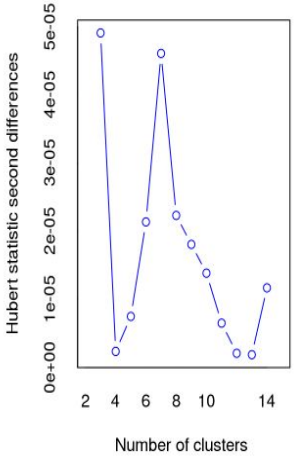


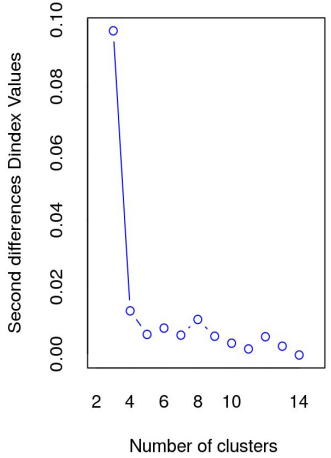
# Description of data preparation

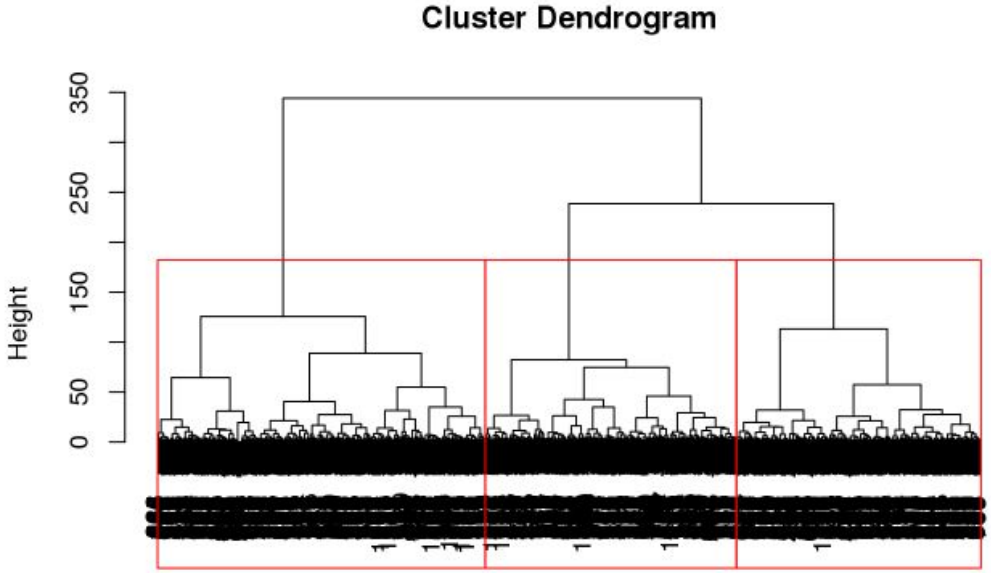
- Data normalisation,
  - All of our textual fields are normalised, since they were used in questions with predefined answers.
- Data subsetting,
  - At this point, we don’t see any need to drop any columns nor rows
- Attribute conversion,
  - Values in 11 qualitative columns will be converted into numerical values in order to perform Ward’s Clustering method


Cluster Dendrogram
Height
0
50
100
200
300
dist(movie, method = "euclidean")
hclust (*, "ward.D")

Based on the graphs below we can notice that three is the most suitable number of clusters proposed by algorithms Dindex and Hubert statistics

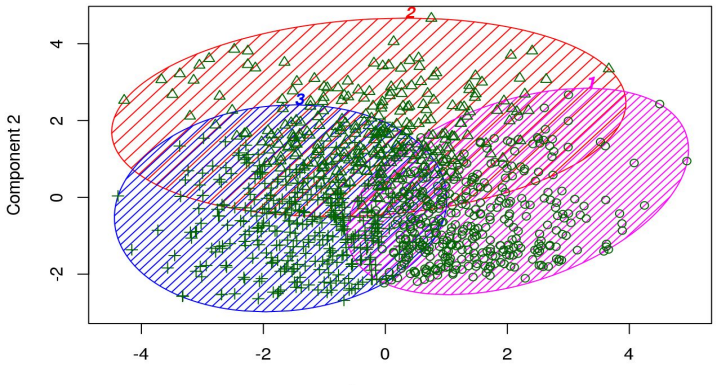


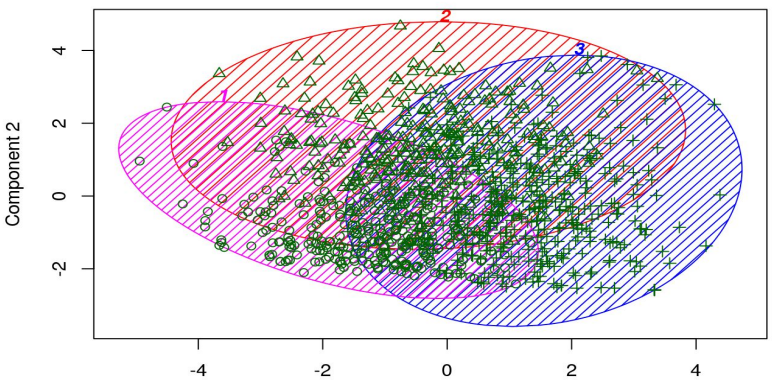




That is why we decided to divide the tree into 3 clusters presented on on the graph

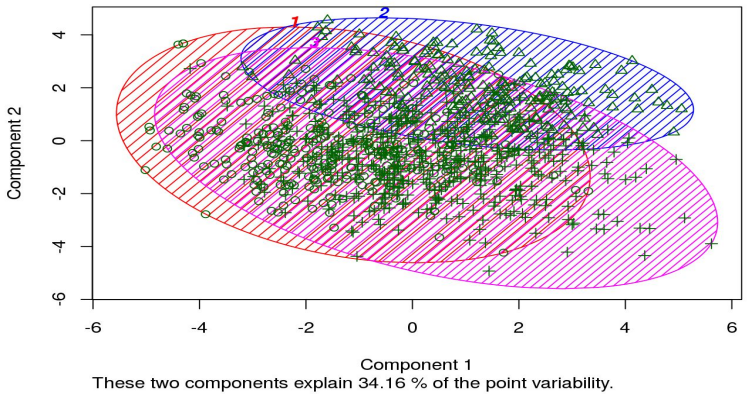
**K-means cluster plot**



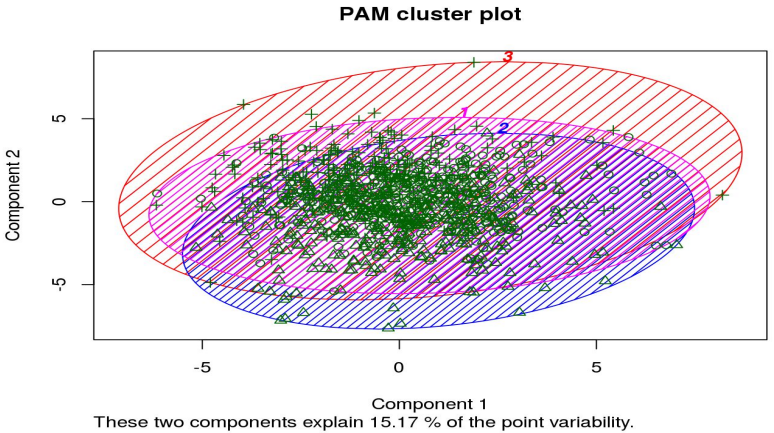

**hclust plot**




**Mclust plot**




**PAM cluster plot**

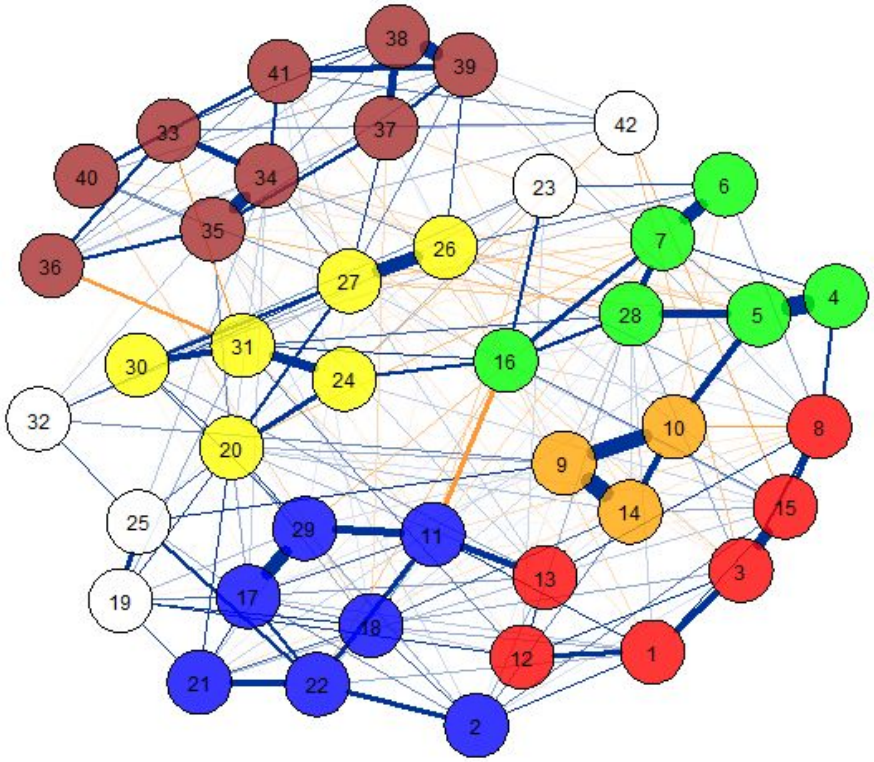


# Cluster summary

- cluster 3, secondary and college male students are the majority
- cluster 1, secondary and college female students are the majority
- cluster 2, master, doctor or another eld youth is the majority, gender is roughly balant in here

K-means has the best cluster performance

Spinglass algorithm




- 1: History
- 3: Politics
- 8: Economy.Management
- 12: Geography
- 13: Foreign.languages
- 15: Law

**B**
- 9: Biology
- 10: Chemistry
- 14: Medicine

**C**
- 19: Countryside..outdoors
- 20: Dancing
- 23: Passive.sport
- 24: Active.sport
- 25: Gardening
- 26: Celebrities
- 27: Shopping
- 30: Fun.with.friends
- 31: Adrenaline.sports
- 32: Pets

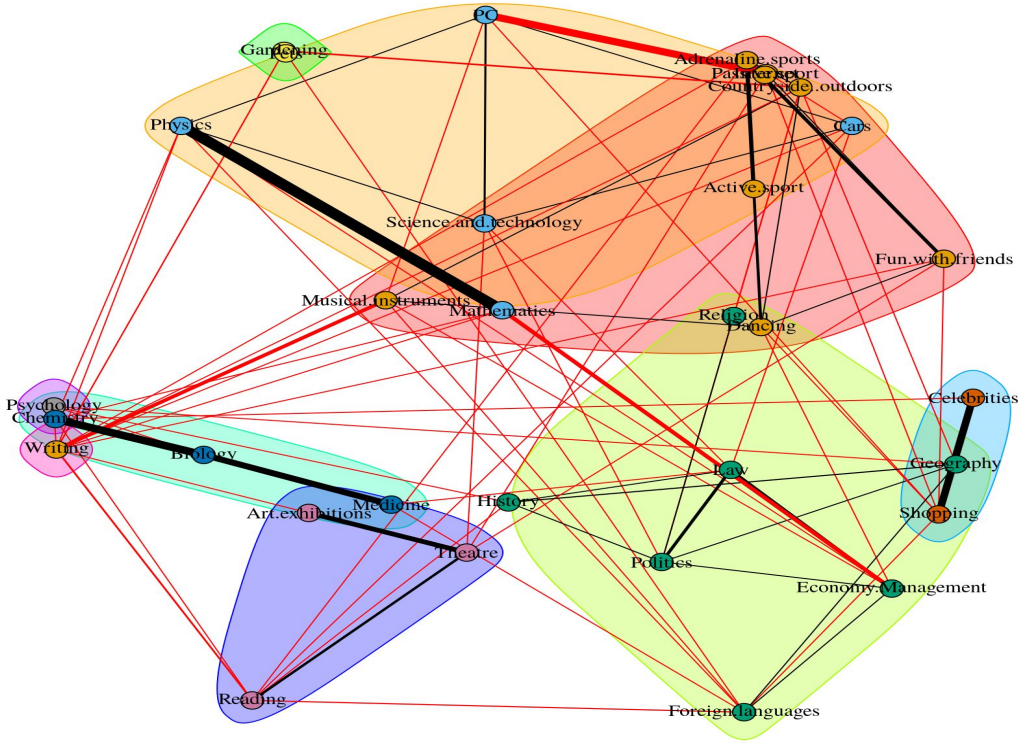
**D**
- 2: Psychology
- 11: Reading
- 17: Art.exhibitions
- 18: Religion
- 21: Musical.instruments
- 22: Writing
- 29: Theatre

**E**
- 4: Mathematics
- 5: Physics
- 6: Internet
- 7: PC
- 16: Cars
- 28: Science.and.technology

**F**
- 20: Dancing
- 24: Active.sport
- 31: Adrenaline.sports
- 30: Fun.with.friends
- 27: Shopping
- 26: Celebrities

**G**
- 36: Heights
- 35: Darkness
- 34: Storm
- 33: Flying
- 40: Ageing
- 37: Spiders
- 41: Dangerous.dogs
- 38: Snakes
- 39: Rats

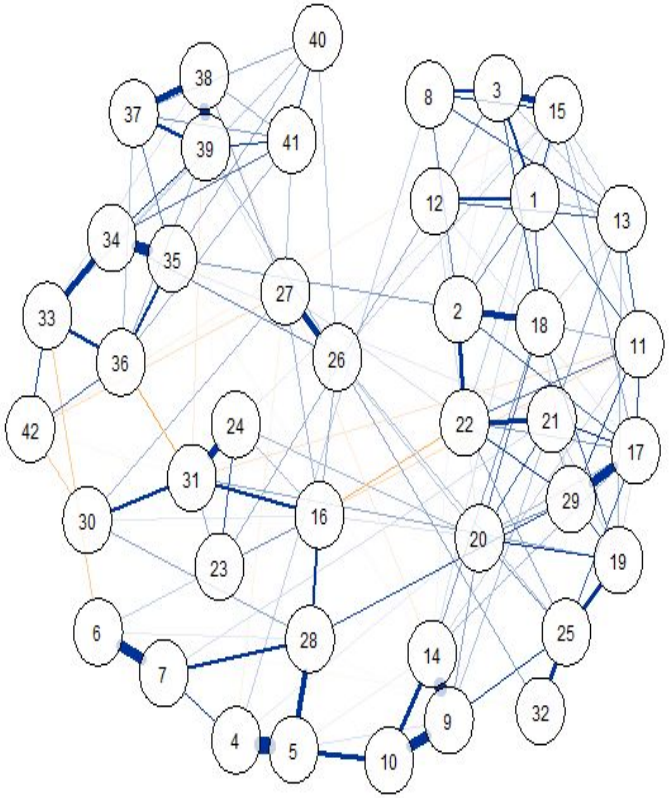
# Walktrap algorithm



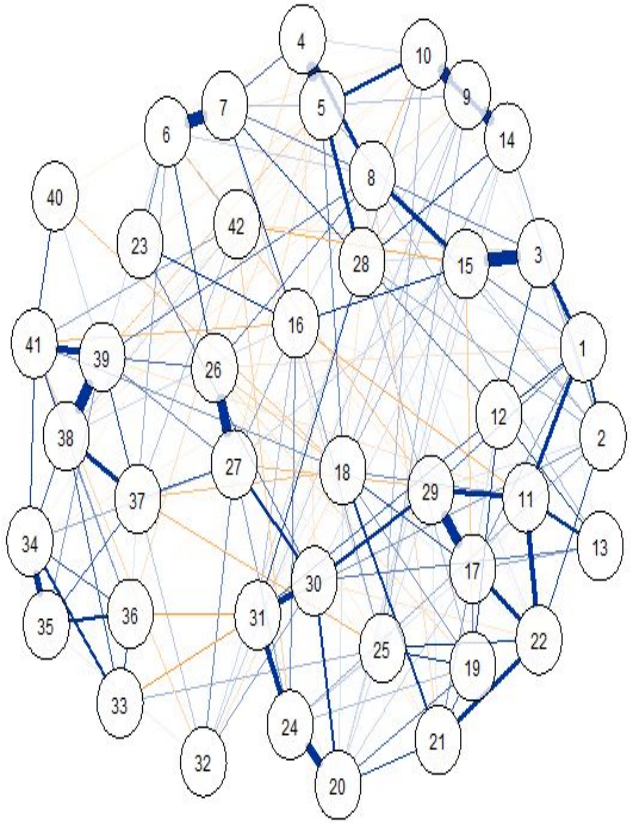
# Network summary

- The Young People who have interest in PC, have also interest in Internet and Learning Science and Technology.
- The Young People who have interest in Mathematics, have also interest in Physics.
- The Young People who have interest in Politics, have also interest in Law and History.
- The Young People who have interest in Reading, have also interest in Theatre, Writing and Foreign Languages.
- The Young People who have interest in Cars, have not interest in Reading and vise versa.

male

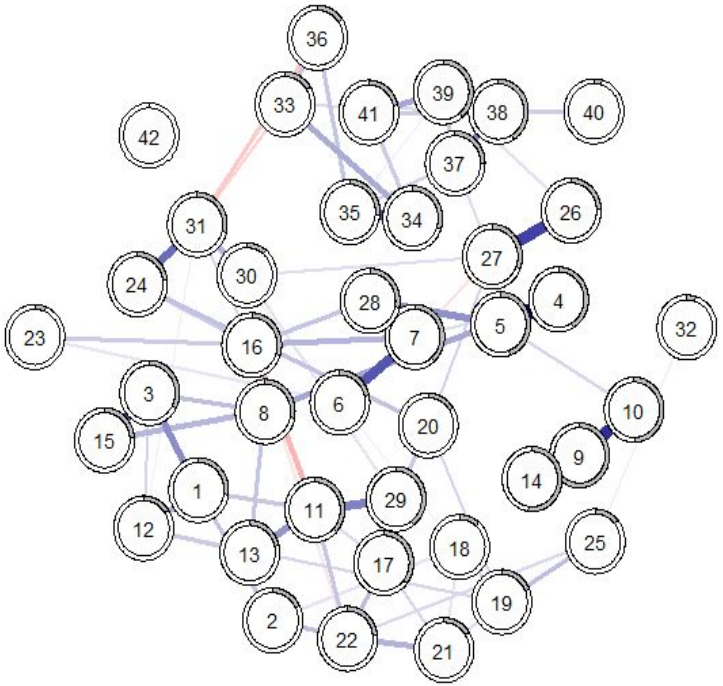



female




1: History
2: Psychology
3: Politics
4: Mathematics
5: Physics
6: Internet
7: PC
8: Economy.Management
9: Biology
10: Chemistry
11: Reading
12: Geography
13: Foreign.languages
14: Medicine
15: Law
16: Cars
17: Art.exhibitions
18: Religion
19: Countryside..outdoors
20: Dancing
21: Musical.instruments
22: Writing
23: Passive.sport
24: Active.sport
25: Gardening
26: Celebrities
27: Shopping
28: Science.and.technology
29: Theatre
30: Fun.with.friends
31: Adrenaline.sports
32: Pets
33: Flying
34: Storm
35: Darkness
36: Heights
37: Spiders
38: Snakes
39: Rats
40: Ageing
41: Dangerous.dogs
42: Fear.of.public.speaking

# Predictability




1: History
2: Psychology
3: Politics
4: Mathematics
5: Physics
6: Internet
7: PC
8: Economy.Management
9: Biology
10: Chemistry
11: Reading
12: Geography
13: Foreign.languages
14: Medicine
15: Law
16: Cars
17: Art.exhibitions
18: Religion
19: Countryside..outdoors
20: Dancing
21: Musical.instruments
22: Writing
23: Passive.sport
24: Active.sport
25: Gardening
26: Celebrities
27: Shopping
28: Science.and.technology
29: Theatre
30: Fun.with.friends
31: Adrenaline.sports
32: Pets
33: Flying
34: Storm
35: Darkness
36: Heights
37: Spiders
38: Snakes
39: Rats
40: Ageing
41: Dangerous.dogs
42: Fear.of.public.speaking

How well can we predict the value nodes by their mutual interactions with neighbouring nodes.