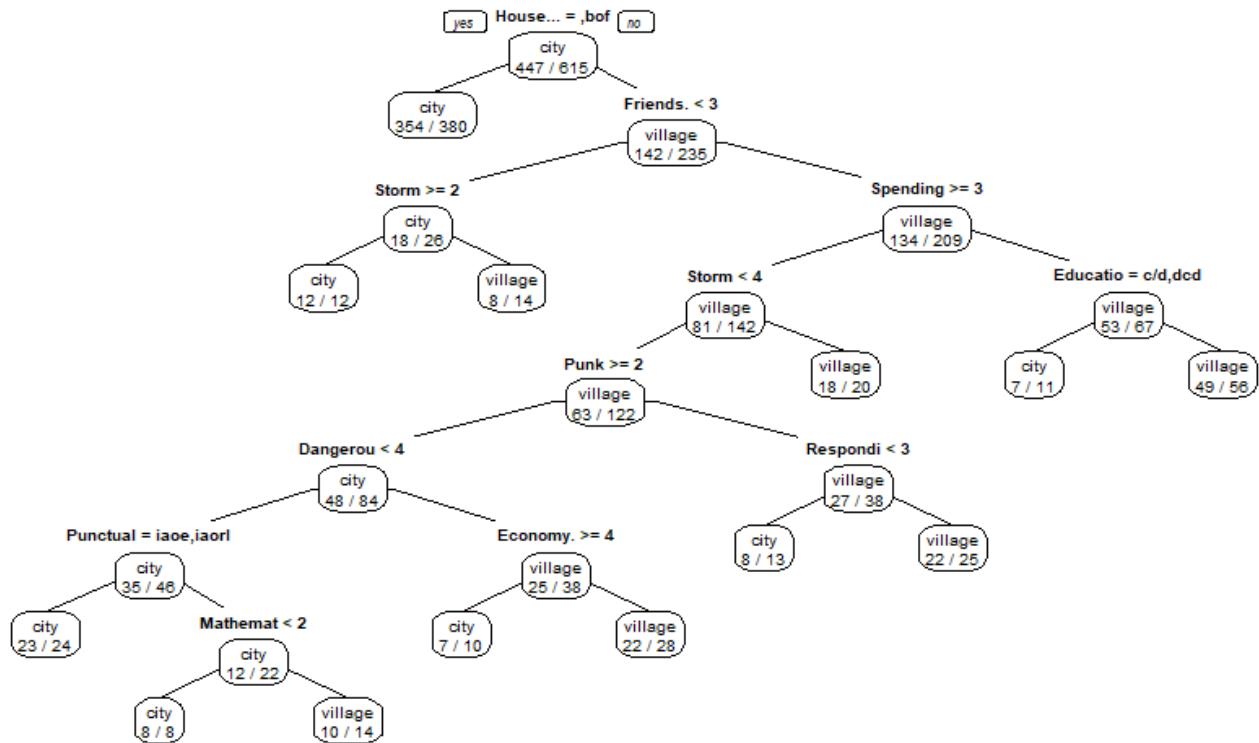


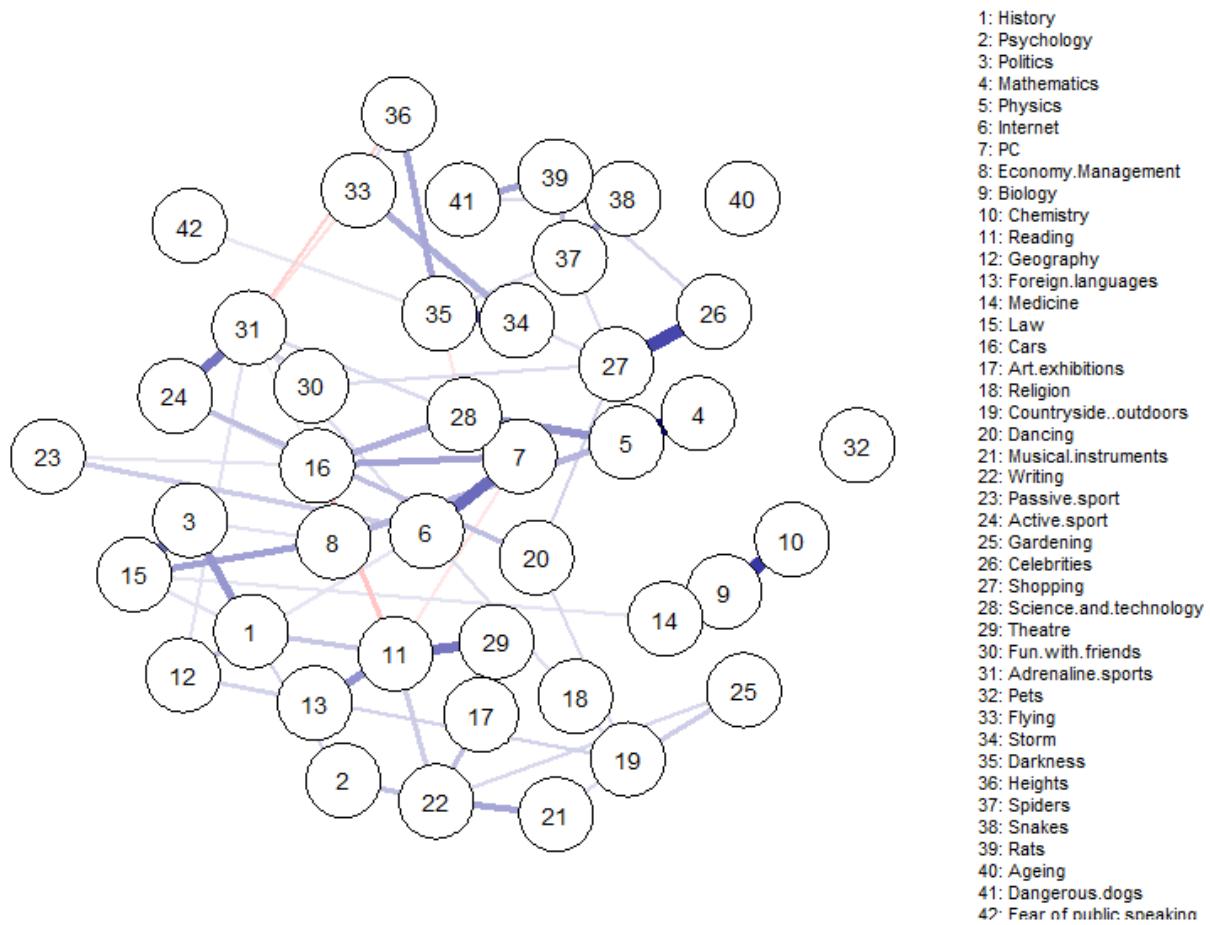
1. Analysis of the obtained results versus the goals that had previously been set

Determining the differences between people with rural vs urban backgrounds



From the chart above it seems that the majority of people, who live in blocks of flats also live in cities. These people are also more afraid of storms and like punctuality. Moreover, it can be seen that people from villages more often say that they have small groups of friends, spend less money on looks, are more into punk music and are more afraid of responding to a serious letter.

Finding the general profile of a ‘money saver’

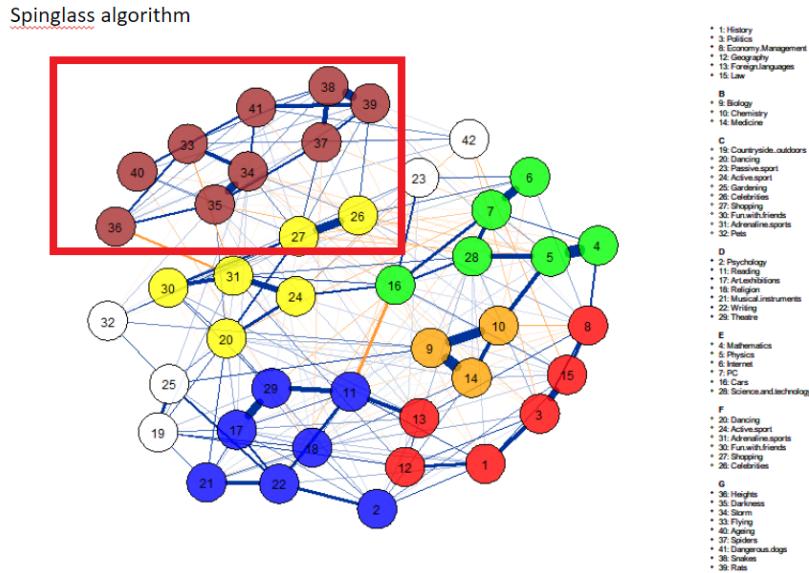


Comparing the diagram of a money saver with the general diagram, we noticed that the main correlation paths are not very different. However, interestingly it seems that money savers have little interest in pets (probably, because they cost a lot of money).

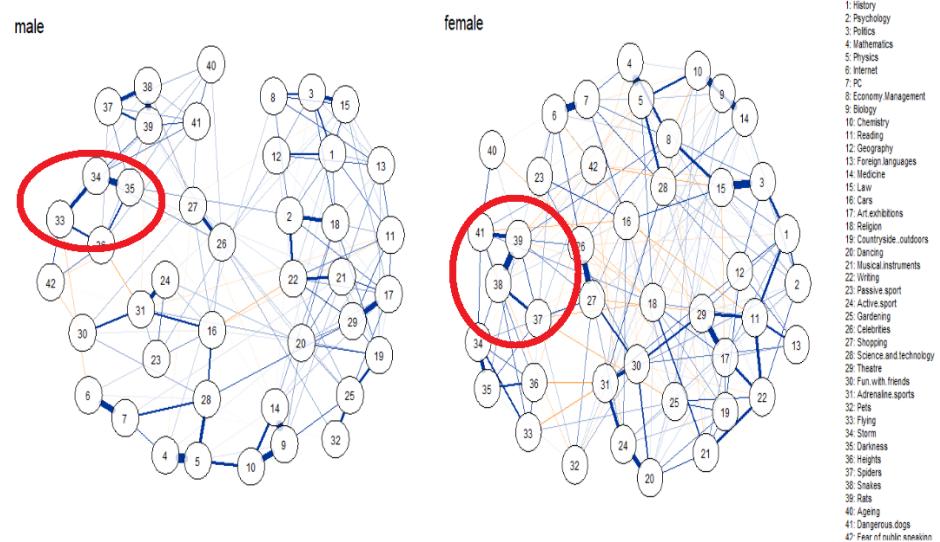
Exploring gender differences in, among others, phobias and fears.

When analysing group G of the Spinglass algorithm for both genders we noticed that there are two groups of highly inter-related phobias

- Flying - Storm - Darkness
- Dangerous dogs - Rats - Spiders - Snakes

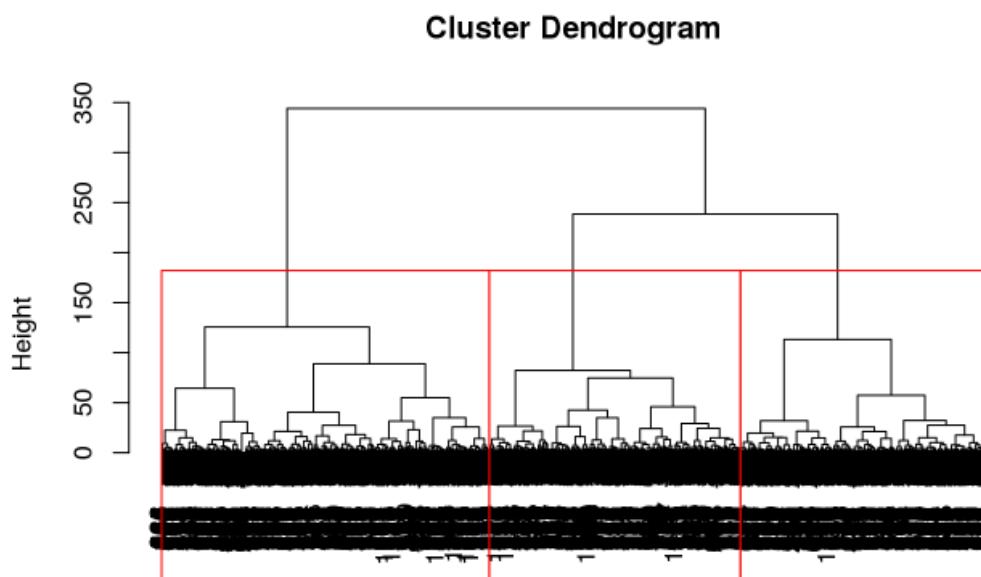


However, by filtering the chart above into the two genders, we were able to see, that each of the above groups is specific to gender. The first group is more prevalent among males, while the animal group is stronger among females.



2. Review of the data mining process

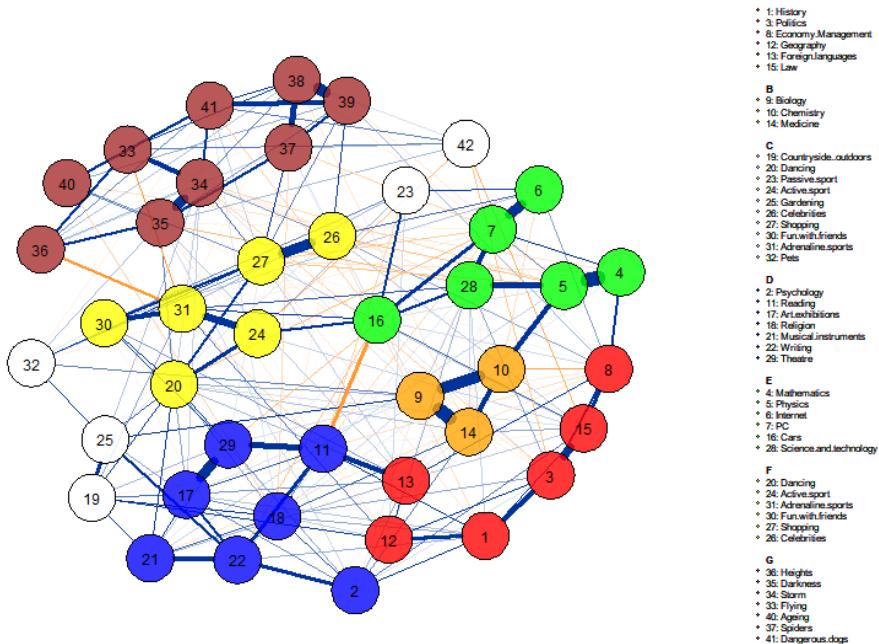
- what has been done (what steps have been executed),
- 1. Data preparation
 - a. Data normalization,
All of our textual fields were normalized since they were used in questions with predefined answers.
 - b. Data subsetting,
We didn't see any need to drop any columns nor rows
 - c. Attribute conversion,
Values in 11 qualitative columns were converted into numerical values in order to perform Ward's Clustering method
- 2. Modelling algorithm
 - a. For the modelling method, we decided to choose a cluster dendrogram



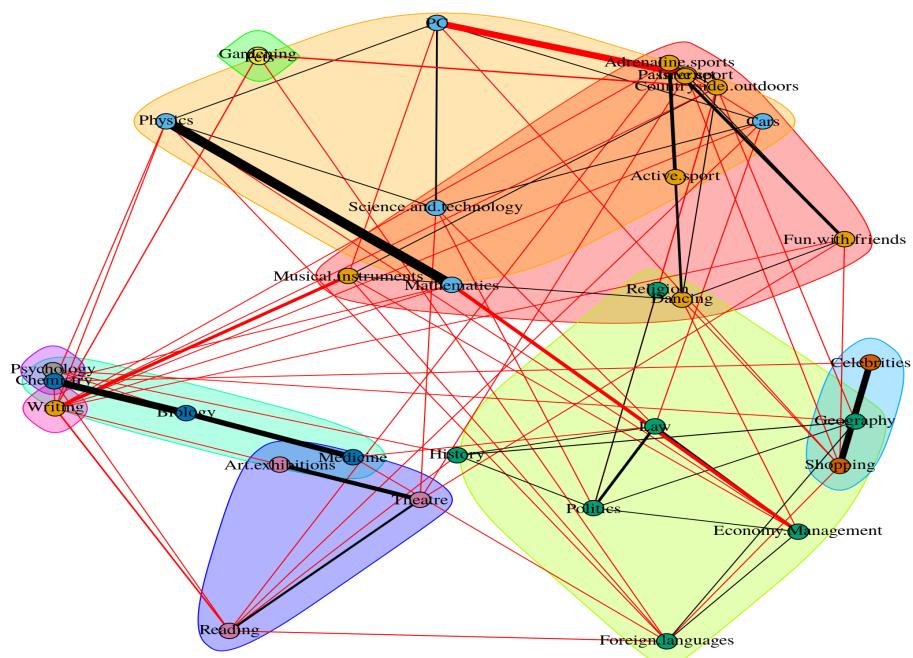
with the number of clusters equal to 3 as suggested by both Dindex and Hubert statistics algorithms.

- b. We validated these clusters using 4 methods:
 - i. Model-Based Clustering and PAM cluster
Both of these methods did not give us any useful insight into the clusters
 - ii. Hierarchical clustering
It was much better than the previously mentioned algorithms with 38,74% of point variability and a much clearer plot
 - iii. K-means (cluster centroid)
This algorithm turned out to give the best results with 38,74% of point variability and the best plot with the least coinciding clusters
- c. For evaluation methods, we decided to use Walktrap and Spinglass algorithms to search for the communities in our data and present the result on network plots.

Spinglass algorithm



Walktrap algorithm



- which steps gave useful results,
The most useful results were obtained with Walktrap and Spinglass algorithm.

The communities were clear and with no errors. We could notice similarities and connections between variables.

For both algorithms, 5 major communities were formed. The walktrap algorithm formed a few smaller communities that we did not notice while using the Spinglass algorithm.

- what could have been possibly done better,

Due to the nature of the network plots, they take longer time than other plots to understand. In future, we should use other plots to present results, divide the data into smaller groups or present the results and smaller and more meaningful networks.

3. Discussion of the importance of the results for further exploration

- what additional data mining tasks could be performed for the dataset, Classification is one of the data mining tasks that could be performed but instead, we decided on doing Cluster analysis. For this particular data set, we could have performed decision trees or random forests to determine the profiles of questionnaire respondents.

- results of which steps of the data mining process could be reused.

Due to the structure of this dataset results of evaluation methods could be reused since it is easy to find communities and correlations on graphs based on networking algorithms. As well as for this data predictability can provide useful results for instance predicting interest in movie categories based on personal beliefs and interest or music taste.