Assignment 6

Part B: Glascow research on smoking habits in teenagers
Adaptive Networks based on the Homophily Principle

The second part of the assignment comprises the run of a contagion model over a data set of teenagers and smoking habits.

All the data was taken from the Description 'Teenage Friends and Lifestyle Study' data.

This folder contains the data and the Python code for the data set provided at http://tinyurl.com/hunndyg.

The purpose of this assignment is to understand how you can model the dynamics of the edges in a social network using the homophily principle. For that, you will have a data set provided[1] and a modeling template based on functions coded in Python that will run the model and relate it to the given data set. (Blankendaal et al, 2016) developed a temporal-causal model for the homophily and more-becomes-more principle combined and tested it using simulations and the same data set you have to use in this assignment, but using the opinion about alcohol drink, instead of tobacco.

Homophily is a principle coined in social sciences that claims that people that share similar traits/opinions/emotions have a higher change of getting a stronger connection. The principle is also known as 'birds of a feather flock together'. The homophily principle can be used to address the strength of the connections in many situations. People with similar political positions tend to be in clusters, as some studies have shown (Conover, 2011; Aral, 2012), and this can be extended to social-economic status, educational level and other traits, besides opinions, feelings and sentiments.

In the data set used, data was collected from 160 students in a window of 3 years. The participants were interviewed once per year, with 3 data points in total. The challenge is to create a model that tries to approximate the evolution over time of both the states and the connections as shown in the data. Remember that in such cases Δt and the speed factors η should have values that provide a good relation to the experiment. So in this assignment, you have to be sensitive to this constraint and try to use the best values trying to fit the model as best as possible.

The matrix of friendships

We have three matrices with the question about the relations in 3 different years.

[1] All the data was taken from "Description 'Teenage Friends and Lifestyle Study'" (https://www.stats.ox.ac.uk/~snijders/siena/Glasgow_data.htm).

The data samples are valued; code 1 stands for "best friend", code 2 for "just a friend", and code 0 for "no friend". Code 10 indicates structural absence of the tie, i.e., at least one of the involved students was not yet part of the school cohort, or had already left the school cohort at the given time point.

For this work, we will make the following changes:

- code 1 = 0.9
- code 2 = 0.5
- code 10 = 0
- code 0 = 0.1

The information about smoking habits

Tobacco use has the scores 1 (non), 2 (occasional) and 3 (regular, i.e. more than once per week). So, for this pattern the values are initialized with values 0.1, 0.5 and 0.9 respectively.

The idea of this code is to simulate the opinion change over time according to the initial values and the network provided by the data set.

Question 1: Why would Δt = 1 year not be a good choice? Why would Δt = 5 months not be a good choice?

Step 0: Documents needed to do this assignment:

- *edges_update.py:* this is where the function to update the edges is. It is here where the homophily effect is coded.
- *states_update.py:* this is where the function to update the nodes (states) is. It is based in a temporal-causal model for the contagion principle as discussed during the course, where a weighted average is used as combination function.
- *data/tobacco.csv:* this comma separated file contains the information about the answers for the questionnaire provided by the students. It presents the 3 data points for each subject for the question about tobacco consumption.
- *data/friendship.1.csv:* this file contains the network of the friendships on the first day. The data are valued; code 1 stands for "best friend", code 2 for "just a friend", and code 0 for "no friend". Code 10 indicates structural absence of the tie, i.e., at least one of the involved students was not yet part of the school cohort, or had already left the school cohort at the given time point. We will also normalize this values to make them be between 0 and 1. The homophily principle will affect this matrix. Note that when a simple linear, quadratic or logistic homophily model is used, due to the factor $W(1-W)$ you cannot work with values 0 or 1 or very close to these values, as no change can occur then. Then it may be more suitable to choose 0.1 as lowest value and 0.9 as highest value for the initial values.
- *Homophily_Glascow_Dataset.ipynb:* this is the notebook (model template) with Python code where you are going to develop the activities. Read carefully the

information at the notebook, and be aware of the images that will be saved in your file.

Step 1: Open the Anaconda interface, and find the folder with the documents for this assignment. Make sure that the libraries needed to run your code are installed (try running the first box).

Step 2: The data set is based on 3 time steps, related to 3 years of questionnaires for the students. That means that we have only one yearly data point. Run the initial part of the notebook until the beginning of step 2 (**STEP 2: Set up**).

Question 2: Choose the parameters for time, Δt and values for the speed factors according to the reality of this experiment and explain your choices, based on the data set.

Step 3: After you run the first time the process, you might change the parameters on step 2 and run again. Adjust them in such a way that the simulation can be more realistic. By realistic it means that the changes are not too abrupt or too slow taking into account the time scale and the number of time steps.

Question 3: What parameters are better for the model? Why?

Question 4: How do you explain the changes in the edges? How does the speed factor affect the steepness of the graphic?

Question 5: What happens if you decrease the amplification parameter (used at the function edges_update)? Why?

Question 6: Is it reasonable to give different speed factors for the update of the edges and for the update of the states of the nodes? Discuss what changes would cause in the model.

Step 4: After you run the model, you should see at the same directory as the Python notebook two files with extension gexf (initial_graph.gexf and final_graph.gexf). Those files are the initial network and the final one. Color the nodes according to their states, and the edges according to their weight using a gradient scale. You should also set the thickness of the edges according to their weight.

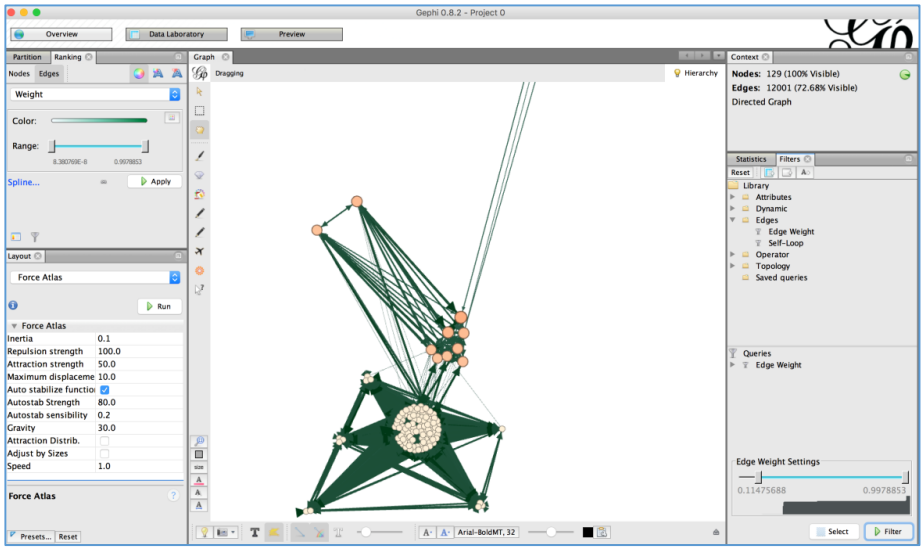
Question 7: Show the plot of the graph after step 4.

Step 5: Try to detect clusters in the network.

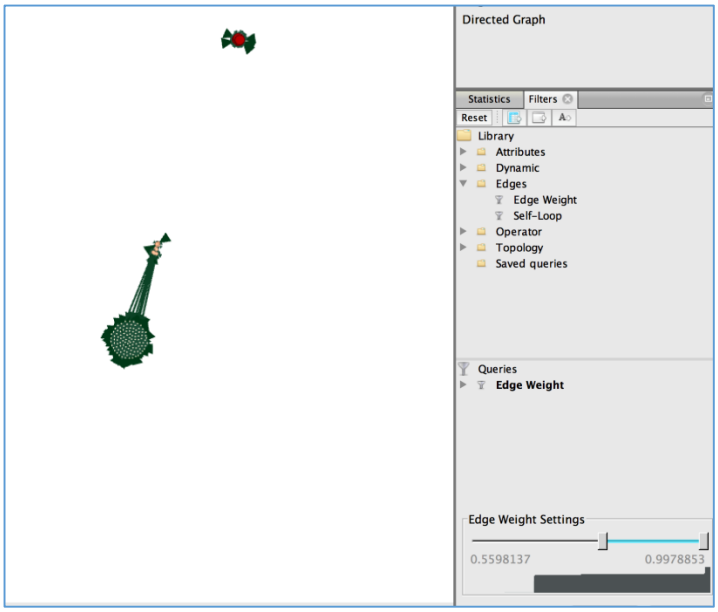
Question 8: Can you explain why these clusters exist? Can you relate the clusters with the graphic of the states plotted before?

Question 9: Make an evaluation of the two networks. Show aspects like centrality of the nodes, and try to find explanations for the changes between the clusters.

Step 7: Filter the edges by their weight. Pick edges with weight bigger than 0.10. Try to find communities in our data set, and bring up characteristics that could explain why a specific cluster exists.



Question 10: Regulate the Edge Weight Settings until your graph is broken in parts. When does that happens? Can you detect important edges that are holding different groups?



Reference

Conover, Michael, et al. "Political polarization on twitter." *ICWSM* 133 (2011): 89-96.

Aral, Sinan, and Dylan Walker. "Identifying influential and susceptible members of social networks." *Science* 337.6092 (2012): 337-341.

Blankendaal, R., Parinussa, S., Treur, J.: A temporal-causal modelling approach to integrated contagion and network change in social networks. In: Proceedings of the 22nd European Conference on Artificial Intelligence, ECAI16 (2016)