HW3

Network Analytics, MSc BA, 2017_18

This is a group HW. It might look long, but I expect you to split the programming part equally amongst yourselves. It will be worth 10% of your grade. We will be covering the material of questions 2 on 11th and 13th of December. Question 3 is econometrics mainly (that I expect you to know from a previous class), but also combined with a diffusion model and some centrality measures we studied. I moved the deadline to 20 December noon.

Submission by **20 Dec 2017 12 noon (this HW will be graded for 60 points)**

Programming: You are allowed to consult on the programming part with your colleagues and on the web but the final code has to be written and debugged entirely on your own.

In the following, you have to search for the appropriate functions in NetworkX/numpy/scipy yourselves.

1. (15 points + 5 challenge)
   The exercise is on detecting communities in two networks
   i. The small Zachary Karate Club network with known community ground-truth
   ii. Your own who-talks-to-whom network (underlying undirected graph where you put an edge if there is a directed link in either direction between two nodes)

For both these networks, you may have to use the networkx functions to go back and forth between edge_list format and adjacency matrix format depending on which community detection package and method you are using. Trying out various graph layouts might help gain some insight.

(a) (do **only one out** of three options)

Option 1: Calculate the community structure using the networkx `girvan_newman` algorithm

Or

(Challenge 5 points)

Option 2: the Louvain `community` package that finds best modularity partition via some fast heuristics. Note that the class data and the village data is in a 0-1 matrix format. You may want to transform it to adjlist format → write to file and →reread, or read each line and add one edge at a time

or

(Challenge 5 points)

Option 3: the igraph package using spectral/eigenvalue methods (these are not available in networkx; installing igraph in python can be a bit of a hassle, but the core is written in C/C++ so it is an order of magnitude faster)

(b) Plot using networkx where different community households have different colors. *(Hints: http://perso.crans.org/aynaud/communities/api.html; http://ryancompton.net/2014/06/16/community-detection-and-colored-plotting-in-networkx/)*

2. (10 points) From the *Bass Model* article in Week 5 folder (first, relate it to the model and terminology we did in class) and the data for the DOCTOR movie (Table 8.7), obtain a rolling-horizon estimate of the parameters (using the optimization model described) for a forecast after observing the sales till week 4. Compare them with the estimates and actual observed demand in the article. You are free to use any regression package and language (R, Excel Data Toolpak, Python).

3. (30 points) In this question your task is to recreate the findings from a network diffusion econometrics study on microfinance (there is a matlab version to help you along). (the NBER working paper in the folder has more details, so that may be more useful instead of the *Science* paper).

The data file is `network_microfinance.zip` in the Week 4 folder. The paper has many estimations. *You have to recreate the first and simpler part really, and all said and done it is just a regression. On the other hand, you will also see network diffusion models in action. I give more tips at the end. The important data is the set of adjacency matrices for the villages (files* `allVillageRelationshipsHH_vilno_XX.csv)`*, the MF dummy on the participation in MicroFinance; cross-check with their calculations in* `cross_sectional.xlsx`

The main network econometrics in the paper has two parts, a cross-sectional estimation and a dynamic structural model part. Your task is to recreate the findings from **only** the first part of the estimation (Do injection points and/or network characteristics matter?).

You may or may not obtain the same results, but they should not be too far off if they are based no similar data. Write (in your own words, 400 words), the model and the conclusions from the analysis. You may want to brush up your notes from your econometrics class.

Tips for Question 3:

1. This model is a regression model rather than the SIR/SIS type models that rely on solving differential equations. However there is an underlying diffusion model that has the same flavor.

The part you have to do is essentially a regression (with 43 data points corresponding to the village files specified in `cross_sectional.xlsx` with corresponding participation data for each village *in MF Dummy.zip) as they do in Section 4.1.*

They relate the fraction who have adapted microfinance (mf) to the network characteristics (such as eigenvector centrality of leader, degree etc).

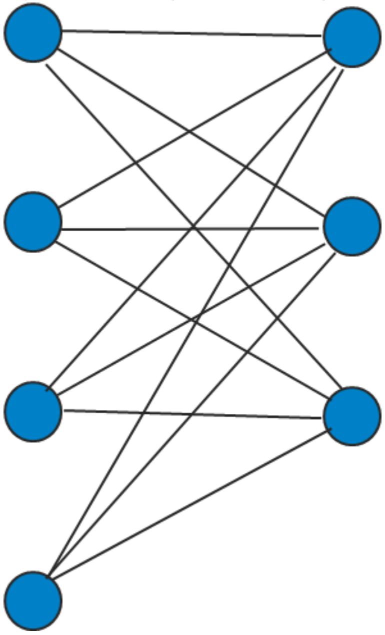
Your task is to (i) get at least the main characteristics out of the network as they do (see cross-sectional.xlsx for what they got out). (ii) run regressions as they do (at least a couple of those in Table 3 of their paper). When I say model, it is the regression model and the usual regression analysis.

2. Load the adjacency matrix for one village. Get centrality measures for one village and then loop through the others once you have it working.
3. There are various adjacency matrices in the data, depending on the relationships. I attach the description of the data sets readme. The URL for the data is here. Use the adjacency matrix `allVillageRelationshipsHH_vilno_XX.csv` which has any edge if there is any one of the types of relationships.
4. To compute the degree of leaders you must compute the degree for leaders who had been surveyed (hhSurveyed==1).
5. The readme says: “Household_characteristics.dta” has demographic information about a household’s home (roof type, number of rooms, latrine type, etc.) and a dummy that indicates whether anyone in the household became a microfinance client but the data itself is sitting somewhere else. I added MF Dummy.zip which contains the dummies for each of the households on whether they participated in the MF.
6. The paper says 43 vs. 77 in the data files. Use the 43 villages in the `cross_sectional.xlsx` rather than the ones in MF Dummy (which has 49 villages) and you’ll get the coefficients close to the ones in Table 3. If you use 49 villages, the coefficient will be far lower, and eigenvector centrality will not be statistically significant.
7. About the network characteristics — I want you to compute them yourself using the Networkx functions.(The “cross_sectional.xlsx” contains their computations. Regressing mf on the other co-variates would be a 10 minute job. This exercise is to learn valuable data munging skills in Python along with network econometrics).
8. The main new work is in merging the MF dummies into the household.xlsx. As the article says, the regression results are in Table 3 and 4 and the co-variates used are there. The main network characteristics are the centrality measures of the leaders. Please ignore the ones not available in the households file. The authors run many different regressions — at least cover the two simplest ones. You can do the regressions themselves in R, Excel or Python.

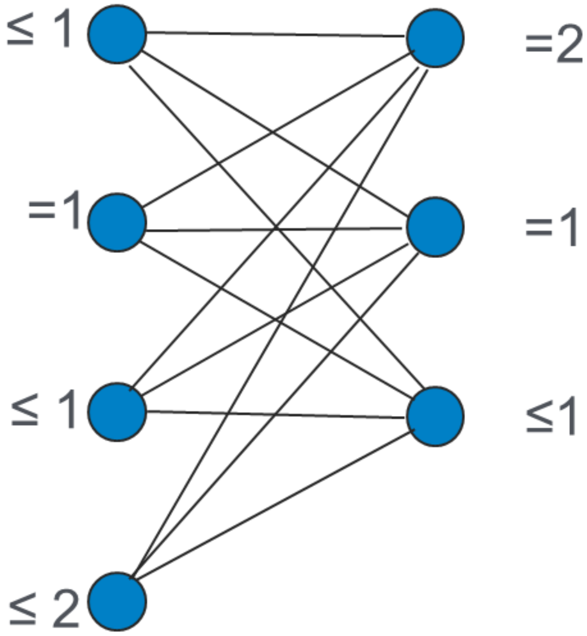
PRACTICE PROBLEMS (Not for submission)

1. Convert the following problems (*Do not give a solution, but a transformation as a weighted perfect-matching problem).*

a. Finding a maximum matching in this graph (note need not be a perfect matching) by converting it to a weighted perfect matching problem



b. Find the largest subset of edges in this graph such that the number of these edges incident on the nodes satisfy the conditions given next to the nodes by converting it to a min-cost flow problem



2. Problem 9, Chapter 10.7 of the EK book