Datasets for Network Analysis Practical Track - 2022

These are three datasets and three scenarios. You can adopt these datasets and scenarios, or chose one of your liking, from other publicly usable files. In that case you need to provide the links to the datasets used.

Deezer:

https://snap.stanford.edu/data/feather-deezer-social.html

Spotify is gaining on us, being able to transform more free users into paying users and we must act. We want to understand our users better and although we have looked into their song preferences through classical machine learning for recommendation systems, we would like now to understand more the social part of our platform, namely the structure of followers. This will allow us to try and adapt products and subscription offerings to different communities. We also suspect that if we tailor some offerings specifically for women, we can get more traction. We heard you are the experts, so if you could let us know:

1- Are there communities in our network, that we can try and understand better? Can you give us information on the top ones?

2- Are there gender specific communities, or differences in the distributions of men and women?

3- Can you give us a list of people that would allow to reach more than 50% of our community?

4 - Any additional information and advice that could help us formulate a strategy would be useful.

Github:

https://snap.stanford.edu/data/github-social.html

We are the marketing department of a software development company that makes software for other developers and companies that develop software for their own needs. Our focus has been pure development, but with the rise of machine learning and their development needs, we would like to expand to this market. When we asked our developers if they could get us some information about the social connections on github (One of the biggest platforms for development purposes, with a social component), they provided us with this information, on users of github and information if they are pure developers or ml dev (machine learning developers). We would like now:

1- To understand if we can directly contact some ml developers to try and have them spread the word? Who to contact then?

2- If there’s a connection between dev and mldev, that we can use to try and follow that route.

3- If the structures of the networks are similar or very different?

4- Any additional information and advice that could help us formulate a strategy would be useful.

githublanguagestats

Enron:

https://snap.stanford.edu/data/email-Enron.html

Our company is growing larger and larger, and the human resources tell us that they fell that communication isn't flowing as expected by email. There seems to be a lot of disconnection and lack of cooperation as is normal as companies grow. They also noted a pattern, where small groups of people would quit at the same time, either to form a company or by being poached by competition, sometimes with serious consequences. They would like to know the seriousness of this problem to help get traction for this initiative, and additional information.

They provided us this dataset of emails of who exchanged emails with whom, during the past few months, excluding HR originated emails. The ID's of the users and any information on the subject or the contents were dropped. We would like to know:

1- How many of our users are disconnected from the main email exchange. And of the main group how many exchange emails with only one other person.
2- On the other side, we want to know if there are many groups forming, and if there are few connections between them. In that case we want to know if there are persons that guarantee our communications between parts of the company, and those that will have an impact in communication spread if they leave.
3- On the other side we want to understand who are the collaborators who maintain good communications throughout the company so we can try and understand if they would be good communication and cooperation ambassadors.
4- Any other information that you think would be helpful would be appreciated.

Although some datasets can be used immediately, some require some work. Some files may need to be opened in a text editor to remove additional information, or to change the file extensions to csv.

Some of the information provided in the dataset files may be useful or not at all. Good judgement will be needed for this work and for future work, use this opportunity to practice it!

Enjoy!

Pedro Alves