CSE5002 Intelligent Data Analysis

Mini Project 2023, SUSTech

Background.

The social network is an essential part in our daily life. Several famous online social network providers such as WeChat and Facebook, heavily rely on the social network analysis so as to make a profit, e.g., via advertising. However, users may be unwilling to provide personal information, which may cause missing attributes in user profiles and thus degrade the performance of the advertising system. For instance, if a product is targeted at users of a certain range of ages, missing ages of some users may lead to inaccurate advertising to them.

In this mini project, an **attributed social network** at MIT, is used as a toy example. The original dataset comes from [1]. To simulate the above scenario, the related term to "age" is "class year" in MIT dataset. Therefore, we adopt "class year" as the label in our mini project. We have preprocessed MIT dataset by removing the lines with 0 presented in "class year", which finally yields 5298 rows of data.

[1] Traud, Amanda L., et al. "Comparing community structure to characteristics in online collegiate social networks." *SIAM review* 53.3 (2011): 526-543.

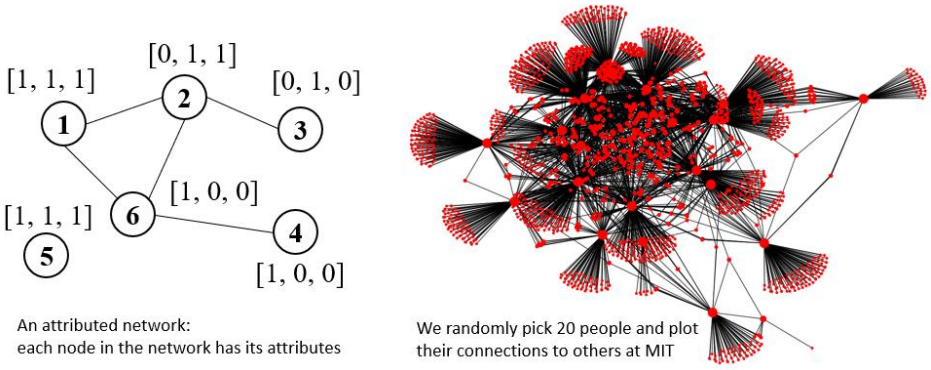

Fig. Left: attribute information. Right: topology information.

Problem Specification.

Assume that there are some missing labels of "class year". We need to predict the missing labels (a multi-class classification problem) based on two sources of information. One comes from node attributes, while another is from network topology. Specifically, our dataset consists of

- **attr.csv**: node_id,degree,gender,major,second_major,dormitory,high_school (5298 rows)
- **adjlist.csv**: node_id,neighbor_id_1,neighbor_id_2,… (5298 rows)
- **label_train.csv**: node_id,class_year (4000 rows)
- **label_test.csv**: node_id,class_year (1298 rows)

where node_id (each corresponds to a person) ranges from 0 to 5297. In this mini project, our **training set** contains node_id from 0 to 3999, and **testing set** contains node_id from 4000 to 5297.

Your objective is to train a classifier, utilizing node attributes, or network topology, or both, to make good predictions for the missing labels in testing set.

Requirements.

- Report.
  - You need to submit a lab report. Some key points are as follows:
    - What is the problem to solve?
    - How do you process the data (including **topology and attributes**) before feeding to a classifier?
    - List all the models of classifiers you have considered and **explain why you choose the final model**.
    - How do you **evaluate** the models?
    - How do you conduct experiments?
    - Please compare and discus your results.
    - What are the limitations and how would you address them in the future?
    - Conduct experiments to discuss whether using both **two sources of information** is better than using a single source of information.
  - Your report is NOT to simply answer these questions. You should organize them properly in a lab report. If you need more advice on the writing, you may follow https://advice.writing.utoronto.ca/types-of-writing/lab-report/
  - Language: English or Chinese in **written language**. Neat typesetting and no grammatical errors are preferred. Hand-writing is forbidden.
- Code.
  - **Runnable** source code.
  - Detailed commands for constructed classes and functions.
  - A **readme file** about how to set up the environment, and how to run your code. Make sure we can run your code properly by following the readme file.

Attention.

- Tools like Chatgpt are not prohibited. If you use Chatgpt, please mark the corresponding part, such as code generation and algorithms' introduction. We encourage deeper thinkings upon the generated content of Chatgpt.
- How to submit. You should compress your **report** (in a pdf) and your **code** (in a folder) into **one zip** file. The zip file with "ID_name", e.g., "10101010_SanZHANG", should be submitted to the link we provided at Blackboard.
- **Plagiarism = 0.** You could discuss with your classmates about the mini project, but please remember not to plagiarize. We will check your report and source code.
- Score: 70 pts (report) + 30 pts (code)