HW4 - Analyzing Bias in Networks

DSCI 531 - Spring 2024 - University of Southern California

Due at 4pm PT Mar 20th, 2024

1 Overview

In this homework, you will conduct basic network analyses on two different networks, perform link prediction on one of them, and analyze bias in networks.

2 Dataset

The friendship networks in UChicago and Caltech. Each node represents a person with a gender 0, 1, or 2: 1 and 2 are two genders, and 0 means gender not specified.

3 Tasks

3.1 Task1 – Network Analysis

1. For each network, calculate the centrality scores of nodes, including PageRank, betweenness centrality, degree centrality, and eigenvector centrality. You can use network analysis tools and libraries for this part, e.g., networkx[1] includes all the needed algorithms. Separate these centrality scores by gender, and compare them. Which network has more gender gap in terms of centrality? Which centrality score(s) show(s) such a gender gap? Give your insights on why this network has a higher gender gap on the centrality score(s).

2. For each network, use Spearman's rank correlation to find the most two similar centrality scores. Why do these two scores have more correlation?

3. For each network, calculate the clustering coefficient of nodes. Calculate the Spearman's rank correlation between the clustering coefficient and the four centrality scores. Which one has the least correlation with the clustering coefficient? Please give your insights.

3.2 Task2 – Link Prediction

Use the Caltech network for this question. We want to perform link prediction on it. In link prediction, we have positive edges and negative edges. Positive edges are edges which are in the graph, and negative edges are edges which are not in the graph. In other words, negative edges are edges in complement of the graph. We train the link prediction model on a fraction of positive edges, and we test the model on how well it can retrieve the rest positive edges.[2]

For evaluation, for each node, we first retrieve the top-k incident edges as ranked by scores given by the model, and then count how many of the retrieved edges are in the test edges, thus obtaining *precision@k* on this node. The *average precision@k* over all the nodes is used to evaluate the model's performance on the entire graph.

[1] https://networkx.org

[2] This is not the only way to for evaluation of link prediction. In some works the model is also evaluated on how badly it can retrieve negative edges, which we will not consider in this assignment.

1. Train-Test split: Use 75% of positive edges (at random) as training edges. Test edges would be the rest 25% of positive edges.
2. Algorithm: Perform link prediction Adamic-Adar and Jaccard Coefficient. The algorithms should output the scores for target edges.
3. Evaluation: Implement the evaluation metric for link prediction based on average precision@k over nodes in the graph. Report the performance on all nodes, on gender1 nodes, and on gender2 nodes. We want to see which algorithm has less bias for genders. On which gender the algorithms give better precision? Which algorithm is more fair?

4 Helpful Resource

- Evaluation Metrics For Information Retrieval

5 Submission Guideline

You will be provided with a Jupyter notebook with detailed instructions for each task. The dataset will also be provided. Please do not use datasets from other sources. Complete the TODOs in the notebook. The notebook and data can be downloaded here. Please include as many comments about your code as possible. You should run every cell and keep the outputs before submitting the notebook.

Please submit your notebook file named as hw4-lastname-firstname.ipynb to Blackboard before 4pm Mar 20th, 2024.

References