PES University, Bengaluru

Department of Computer Science & Engineering

Session : Jan-May, 2022

UE19CS345 – NETWORK ANALYSIS AND MINING

Lab Evaluation 02

Submission Date: 29th April 2022, 23:59

TITLE: Comparative Analysis of Traditional Methods and Graph Machine Learning Methods for Link Prediction

GOAL

To compare traditional, similarity-based methods with GraphML algorithms like GCN, GraphSAGE and GAT for link prediction.

PROBLEM STATEMENT

Every team has been assigned one of eight datasets randomly. Each of these datasets represents corresponding seasons from the TV Show, Game of Thrones, where each dataset consists of two CSV files - nodes.csv and edges.csv The former provides the list of characters whereas the latter provides the interactions between them. This is similar to what was extracted in Lab Evaluation 1. (The GitHub repository for the datasets can be found here.)

The aim is to perform link prediction on this dataset using both traditional and graph machine learning. **In this assignment, we will treat link prediction as a binary classification task (i.e. 1 implies that there is a link between the nodes whereas 0 denotes the absence of a link).** Subsequently, the methods and thresholds need to be compared and analysed. Any insight drawn from the modelling that matches the storyline should also be stated.

REQUIREMENTS

1. Load the dataset from the CSVs into NetworkX to create an undirected graph.
2. Perform EDA on the dataset
    a. Check for isolates, self-loops, etc
    b. Can be minimal (refer to Lab Evaluation 1 to see what kind of EDA can be done)
3. Calculate the following measures:
    a. Betweenness

b. PageRank
c. Local clustering coefficient

These would be the node features in your graph. You are also encouraged to use other node features in addition to these with proper justification as to why they were used, if they help with providing better metrics.

4. Find communities using Spectral Clustering - Code (for reference only) can be found here.
    a. Provide hypotheses about how these communities might be getting detected, according to the plot of the season (for example: are they being formed based on the houses the characters belong to?)
5. Link Prediction using Traditional Methods - Code (for reference only) can be found here. It is recommended that the proximity-based likelihood approach be used, but feel free to use the top-k approach if need be.
    a. Perform comparison for **at least** 3 similarity measures - we recommend using Jaccard, Adamic Adar and Preferential Attachment, but you are free to try any others in addition to this.
    b. Find the best similarity measure along with the optimal threshold for said similarity measure.
6. Link Prediction using GraphML - Refer to the SEAL paper.
    a. Load the graph from NetworkX into PyTorch Geometric
    b. Perform link prediction using
        i. GCN
        ii. GraphSAGE
        iii. GAT
    c. Compare the three models - **use Loss and AUC for comparison** (this holds even for traditional link prediction)
7. Perform a comparison between traditional and GraphML
    a. What are the metrics?

b. Which is performing better? Is there any reason you can think of, as to why this might be happening?

c. Any analysis or insights you can draw from this, that may relate to the season's plot?

MARKING SCHEME

1. EDA and Community Detection using Spectral Clustering
    a. Implementation - 1 mark
    b. Analysis - 1 mark
2. Traditional Link Prediction
    a. Implementation - 2 marks
    b. Analysis - 1 mark
3. GraphML
    a. Implementation - 3 marks
    b. Analysis - 1 mark
4. Comparative Analysis - 1 mark

IMPORTANT POINTS TO NOTE:

1. The type of similarity measures, the architecture and the design of the GraphML models are up to the team's discretion. However, we will expect justification for every design decision taken by the team in the form of documentation.
2. You are **not allowed** to use context from the books or any other season apart from the one assigned to your team to make design decisions. Please restrict your implementation and your analysis to the season assigned to your team. If any other dataset apart from the one assigned is used, a penalty will be imposed.
3. There is no separate analysis section in this assignment. Therefore, we require you to appropriately document your code using the Markdown feature wherever necessary. Any insights that are drawn should not be mentioned at the end but rather, in line with the code that justifies it.