3. Analyzing citation networks

This task involves exploring the High-energy physics citation network. Arxiv HEP-PH (high energy physics phenomenology) citation graph is from the e-print arXiv and covers all the citations within a dataset of 34,546 papers with 421,578 edges. If a paper i cites paper j, the graph contains a directed edge from i to j. If a paper cites, or is cited by, a paper outside the dataset, the graph does not contain any information about this. This dataset is temporal, which means the structure of the network changes over time as new academic papers are published.

Data: http://snap.stanford.edu/data/cit-HepPh.html

a. Task 1 : The first task is a graph exploration task. Build out a graph from the dataset given, and record how the graph and its properties change over time. You are expected to perform this task on at least 5 properties, and report interesting insights. Few simple properties include different types of centrality, density, and diameter. This task is focused on exploratory data analysis, and you are expected to show plots and metrics to support your findings.

b. Task 2: Community detection or clustering is an important analysis for graphs. In the study of complex networks, a network is said to have community structure if the nodes of the network can be easily grouped into disjoint sets of nodes such that each set of nodes is densely connected internally, and sparsely between different communities. In this task, you are required to perform community detection on the graph. This is a well studied problem, and various static algorithms as well as machine learning methods exist for community detection. You are required to:
    1. Implement any two algorithms/ ML methods for community detection on the graph at any time T
    2. Analyze the communities (Can you build an understanding of why the algorithm chose the communities it did?)
    3. Perform temporal community detection, through which you can study how communities evolve over time as new papers are added. Report interesting insights using various plots and metrics

c. Bonus Task []: Link Prediction is a task in graph and network analysis where the goal is to predict missing or future connections between nodes in a network. As before, multiple algorithms exist for this task. However, you are required to implement both a graph neural network, and a classic algorithm like DeepWalk or Node2Vec. For training, you can use the citation network at any time interval [0-T] and for testing/validation, use nodes that appear after time T. ,. Compare the results of the two approaches, and analyze whether the GNN performs better, and if so, why. You will be evaluated on how well the model can predict these edges, as well as your understanding of the link prediction task in graphs

d. Paper Reading Task : https://arxiv.org/pdf/1607.00653.pdf

Resources:

- Networkx Python Library
- https://pytorch-geometric.readthedocs.io/en/latest/index.html
- https://web.stanford.edu/class/cs224w/