Network Analysis of Massachusetts Institute of Technology Scholarly Work

Motunrayo Ibiyo

Problem Statement

It is estimated that at least 64 million academic papers have been published since the year 1996, with the growth rate of newly published articles increasing over time. As of 2022, over 5.14 million academic articles are published per year, including short surveys, reviews, and conference proceedings. A key aspect of academics' paper is the paper that is cited and the impact of the paper's future publication. This creates a form of relationship among academic publications. This characteristic of this relationship poses to be complex given the many to many relationships that exists among publications over time.

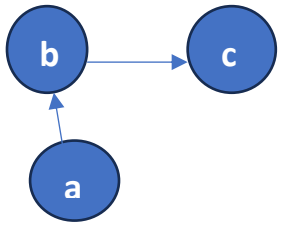
Problem Importance

Considering the magnitude of publication that is in the world, it is essential to understand the characteristics that are inherent in the process authors of publications choose to decide on the papers to present as reference. Citation analysis is a method that has been used in the past to gauges a publication for its trustworthiness. This project aims to find out attributes that contribute to the trustworthiness of a citation.

Proposed Approach

To conduct a citation analysis on citation, the following method will be used:

1. Download "MIT Scholarly Works Over Time" dataset from I³ Open Innovation Dataset Index. I³ Open Innovation Dataset Index is a publicly maintained index of datasets relating to innovations and patents. The MIT scholarly dataset is maintained by The Lens. The dataset contains sscholarly works produced by MIT 1950-2018.
2. Extract the source and target node from the dataset. The project is aimed on citation relationship between papers so the network will be a directed graph like the one below. Paper "a" cites paper "b" and paper "b" cites paper "c"



3. The extracted node will be saved into a file.
4. The extracted nodes will be analyzed using networkx python library.

Expected Outcome

The analysis will aim to answer the following questions:

1. What is the most popular publication in this dataset.
2. What is the average number of citations a paper has.
3. What is the relationship between publication type (Journal paper, conference paper etc.) and citation popularity.
4. What is the relationship between paper publisher and citation popularity.