Comparative analysis of structural, learned, and hybrid feature representations for graph clustering

Learning From Networks - Project Proposal

Leonardo Gusson, Luca Rao, Chiara Frizzarin

1 Motivation

This project aims at solving a question: is it possible to use a co-purchasing network to find out the actual categories of products considered? And which kind of features representation works better: structural, learned or hybrid?

The analysis is based on the "Amazon product co-purchasing network metadata" which was collected back in 2006 by crawling Amazon website.

From this dataset a directed graph $G = (V, A)$ is going to be built:

- **Nodes**: Each node $v \in V$ represents a unique product in the Amazon dataset ($|V| = 548,552$);
- **Arcs**: An edge $(u, v) \in A$ exists if product v is often co-purchased after u ($|A| = 1,788,725$).

Each node comes with a set of information following this format:

- **Id**: Product id (number 0, ..., 548551)
- **ASIN**: Amazon Standard Identification Number
- **title**: Name/title of the product
- **group**: Product group (Book, DVD, Video or Music)
- **salesrank**: Amazon Salesrank
- **similar**: number $n \in [0, 5]$ of co-purchased products followed by a list of their ASINs (people who buy X also buy Y) (e.g. 2 B0001500VS B000002WA3)
- **categories**: Location in product category hierarchy to which the product belongs (separated by |, category id in [])
- **reviews**: Product review information: time, user id, rating, total number of votes on the review, total number of helpfulness votes (how many people found the review to be helpful)

> A deeper analysis of the directed graph $G = (V, A)$ built from the Amazon co-purchasing dataset pointed out that weakly connected components are significantly heterogeneous. Specifically, a main connected component was identified, containing 334,843 nodes out of a total of 542,664 valid nodes. The subsequent connected components were found to be negligible in size (for instance, the second and third components contained only 222 and 184 nodes, respectively). Therefore, for this problem we will use the largest connected component of the graph to simplify the implementation of the algorithms.
>
> co-purchasing graph pre-processing: We decided to retain only products belonging to the categories 'Book', 'DVD', 'Video', and 'Music', along with their ASIN, title, group, and salesrank attributes. Subsequently, the largest connected component of the resulting directed graph was extracted and saved as a serialized NetworkX object in the .pickle format (chosen for efficiency) for downstream analysis.

All this information and the dataset can be found at Stanford Large Network Dataset Collection.

2 Method

The objective of this project is to compare the effectiveness and performance of three