Graphing Substitute-Complement Relationships

Omar Abdel Haq

APMTH 220 Poster Session, Harvard University, Cambridge, Massachusetts 02138

Abstract

This project investigates the efficacy of Graph Neural Networks (GNNs) in identifying substitute and complement relationships among products using Amazon's co-viewing and co-purchase data. By leveraging GNNs to model product interactions, we aim to discern patterns indicative of substitute or complementary goods, providing valuable insights into consumer behavior and market dynamics. Preliminary results demonstrate the robustness of GNNs to noise and their potential in capturing nuanced product relationships, paving the way for more accurate economic analyses and predictions.

Introduction

This project explores the potential of using Graph Neural Networks (GNNs) for detecting substitute or complement relationships in commerce. Understanding whether products are substitutes or complements is crucial for economists because it helps in analyzing consumer behavior, market dynamics, and making predictions about how changes in prices or demand for one product might affect another. However, no large scale methods exist to detect common complementary or substitute products.

The overarching project motivation is to determine whether online-purchase data obtained from Amazon can be used to detect with high confidence substitute or complement pairs; for this we use a link prediction approach using GNNs on co-viewing and co-purchase Amazon data. Link prediction algorithms have been utilized for tasks such as suggesting friends on social media platforms, recommending products to users based on their preferences, and predicting future interactions in social networks, but no significant body of literature exists on their utility in evaluating the socioeconomic qualities of goods.

Background

This project attempts to model complementary and substitute goods using the Also-View and Also-Buy features of Amazon products, which indicate whether customers viewed sets of products at the same or purchased them together, respectively. Complementary goods are defined as products that are typically used and/or purchased together, while supplementary goods are alternatives that can substitute for each other, with consumers usually picking between them when making purchases.

The key assumption we make is that goods that are viewed together but not purchased together (with only one of them bought) are likely substitutes. On the other hand, goods that are viewed together and bought together are likely complements. This however, is not always strictly true; some users buy similar items and only keep the one they like, while some pairs of goods are neither substitutes nor complements but are still purchased together. We use GNN modeling to predict co-viewing and co-buying, in hopes of filtering out those noisy cases.

Approach

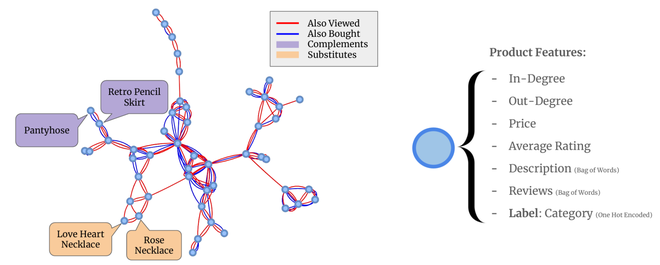


Figure 1. *Sample node relationships & node feature structure.*

This project uses the Amazon Review Data, a dataset compiled by researchers at UCSD. This dataset offers a comprehensive collection of user reviews, product metadata, and connections, offering valuable insights into consumer behavior and product characteristics within the Amazon platform. Two homogeneous graphs were created where each node represents products; each node contains information on the product's price, rating, and bag-of-words vectors of the product's description and reviews left by buyers. One graph's edges represent co-viewing status while the other's represent co-purchasing status. The labels are one-hot-encoded category vectors of each item.

Nodes	Features	Also-View Edges	Also-Buy Edges	Categories
58,779	2,674	97,491	101,336	3

Table 1. *Dataset attributes.*

The experimental approach is to run different variants of GNNs to create embeddings for the product nodes, and then apply another network to calculate the probability of two nodes being connected by a link. We withhold 5,000 of the dataset's edges for validation and another 5,000 for the test set. Negative sampling of fake edges is used so that the link prediction network doesn't naively assume any two nodes should share a link (the case if we only have positive edges in the dataset).

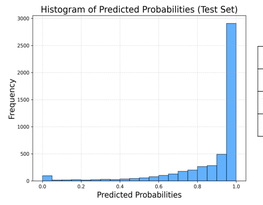
We train a convolutional GNN to create embeddings either using the GCNConv or SAGEConv layers for both link prediction tasks and with varying levels of noise. Noise is defined as negative-sampled edges that we make the model assume are positive edges. If modeling allows us to arrive at a model with high robustness to noise, then the link predictor is a lot likelier to capture true substitute or complement relationships.

Results

Layer Type	Noise	Buy, Train	Buy, Test	View, Train	View, Test
GCNConv	0%	97.66%	95.44%	97.96%	**96.28%**
GCNConv	5%	97.89%	**95.72%**	97.25%	95.30%
SAGEConv	0%	97.80%	93.54%	97.45%	93.66%
SAGEConv	5%	96.25%	91.64%	97.55%	94.30%

Table 2. *Preliminary model results.*

From the preliminary results table above, we see that the addition of fake/noisy edges during the training process does not substantially worsen the performance of our models, and in some cases it actually improves their performance when compared to a no artificial noise scenario.



Product A	Product B	Also-Buy Prediction	Also-View Prediction
Gold-Tone Watch	Chronograph Watch	37.21%	99.32%
Ultra-Violet Flashlight	Moonstone Bracelet	3.46%	9.80%
Santa Claus Beard Set	Granny Glasses	99.48%	99.10%

Figure 2. *Link predictor predicted probability distribution & sample observations.*

The link-predictor (selected at random from the ones trained above) does not consistently output high probabilities for all positive edge cases, in many cases output lower link probabilities. By inspecting sample nodes that do have edges in the test set (represented in the table above) we see that the predictor does not consistently predict a link.

Discussion

From the results above, we see that a GNN approach to create node embeddings does seem to be robust to noise, both artificial and real. The link-predictor used for the visualizations outputs low probabilities for pairs of products that have little relevance to one another despite having a link, and performs well (sometimes even better) when compared to its noiseless counterpart. *Its performance indicates an ability to capture substitute & complement relationships*.

Future Direction

A potential avenue for further work is creating larger-scale visualizations of the model's outputted relationship predictions, as well as looking into other types of product relationships that Amazon provides (like "Buy After Viewing").