CS4195 Modeling & Data Analysis in Complex Networks Final Assignment - Group11

Berend Baas
Delft University of Technology
Delft, South Holland
email@student.tudelft.nl

Jody Liu
Delft University of Technology
Delft, South Holland
email@student.tudelft.nl

Kyriakos Psarakis
Delft University of Technology
Delft, South Holland
email@student.tudelft.nl

Panagiotis Soilis
Delft University of Technology
Delft, South Holland
p.soilis@student.tudelft.nl

1 INTRODUCTION

This assignment focuses on analyzing the co-purchasing behaviour of Amazon users. In particular, this is done by creating a network of Amazon products where the edges connect products that have been co-purchased. The assumption made in this case is that more product reviews correspond to more product purchases. Therefore, we further evaluated the network to reason about the most influential products and whether the review distribution can be approximated via our proposed network. The code implemented for this study is publicly available on GitHub[1].

2 DATASET

The dataset selected for this assignment is the Amazon product review dataset provided by Julian McAuley, UCSD [2]. To be more specific, it was originally created by [1]. The original dataset contains 548,552 products from the categories of Books, music CDs, DVDs and VHS video tapes, where all products have 0-5 recommendations. Due to the creation date of the dataset we filtered products which have at least two years of reviews and extracted the review count from the last available year. Following that we randomly sampled 10,010 nodes from the largest sub-graph.

3 METHODOLOGY

The co-purchase behavior can be analyzed by creating a network where the products are the nodes and the edges indicate that two products are co-purchased. This assignment is divided into two parts:

Review count approximation

We tried to approximate the popularity of each product by using a Random Walk and a Generalized Random Walk, and calculated the normalized visits of each product within the walk. The visiting probabilities for each iteration were calculated as follows:

- *Random Walk*: $\pi_i = \frac{1}{d_j}$
- *Generalized Random Walk*: $\pi_{ij}^t = \frac{v_i^t + 1}{\sum_{b \in Nodes} a_{jb} v_b^t + d_j}$

where v_i^t corresponds to the number of visits that node i has at timestep t, d_j is the degree of node j and a_{jb} is the adjacency matrix value for jb. The output of the two walks was then compared to the true popularity distribution which is based on the number of reviews for each product.

Influential nodes

The second part of the assignment focuses on quantifying the influence of removing nodes with different characteristics from the network. To that end, three different networks were created:

- *Top 1%*: top 100 nodes with highest popularity get removed from the node
- *Bottom 1%*: bottom 100 nodes get removed
- *Random 1%*: 100 random nodes are dropped from the network

Following that, the three networks are compared to the original one using the l1-norm to quantify their "distance" from the original network. This procedure was implemented for both the Random Walk and the Generalized Random Walk.

4 EXPERIMENTS

The aforementioned methods were tested by running the Random Walk for 10,000,000 iterations and the Generalized Random Walk with 1,000 random initial points using 10,000 iterations for each one. The experiments performed resulted in three popularity distributions which can be found in Figures 1, 2 and 3 respectively.

Concerning the influence nodes test, the values obtained can be found in Table 1.

l1-norm Distance	Random Walk	Generalized Random Walk
Top 1%	0.2668	-
Base 1%	0.0665	1.8534
Bottom 1%	0.0583	1.7968

Table 1: Node removal l1-norm

[1] https://github.com/kPsarakis/Complex-Networks-Final-Project-Group11
[2] http://snap.stanford.edu/data/amazon-meta.html

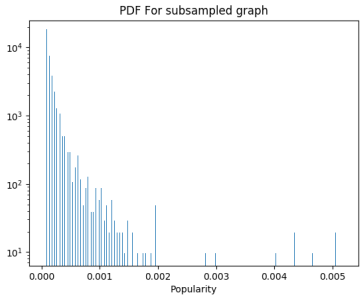

Figure 1: True Distribution

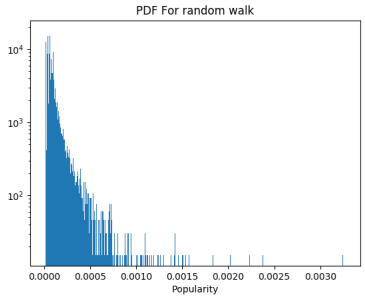

Figure 2: Random Walk

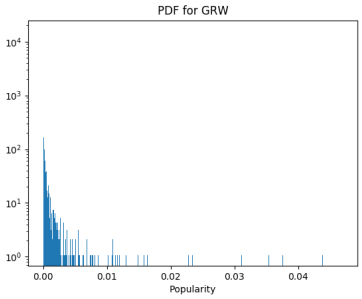

Figure 3: Generalized Random Walk

The findings arising from our research are the following:

- The Random Walk method better approximates the true product popularity.
- The standard version of the Generalized Random Walk "overfits" on the node visits. The use of a hyperparameter α that puts less weight on the node visit history could potentially solve the issue.
- The removal of the top 1% has a much larger influence in the network structure than the random 1% or the bottom 1%. Moreover, while the random 1% has a higher influence than the bottom 1%, their values are very closely matched.

REFERENCES

[1] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.