

Abstract

Amazon provides a particular feature as part of their recommendation systems called “Customers Who Bought This Item Also Bought...” which uses predictive analytics in order to present products that have a high probability of being an interest to the user.

This report aims to study the amazon co-purchasing network whereby finding the relationship between purchased products and the products that were purchased as a result of purchasing it.

From the investigations, it was found that the in-degree distribution displayed a power law relationship but the out-degree didn't show a power law relationship.

It was also found that there is no rich club structure within the network and that the network was found to be moving away from being assortative i.e. products being purchased with products that have similar degree.

The investigations did yield an interesting result which was that there is a narrow range of products that could potentially be co-purchased for a given product.

CONTENTS

1	Introduction.....	3
2	Background.....	3
2.1	Degree distribution	3
2.2	Assortative coefficient ^[7]	3
2.3	Rich club coefficient ^[7]	4
3	Literature Survey	4
3.1	Analysis of product purchase patterns in a co-purchase network ^[8]	4
3.2	Prediction of Co-purchasing products ^[9]	5
3.3	Motif Analysis in the Amazon Product Co-Purchasing Network ^[10]	5
4	Methodology	7
4.1	Assortative coefficient algorithm	7
4.2	Rich club coefficient algorithm	7
5	Results	8
5.1	Basic features.....	8
5.2	Degree distribution	9
5.3	Assortative coefficient & rich club coefficient.....	11
6	Discussion	12
	References	12

1 INTRODUCTION

Amazon provides a particular feature as part of their recommendation systems namely “Customers Who Bought This Item Also Bought...”. This feature incorporates predictive analytics to present products that have a high probability of being an interest to the user.

There are two ways recommendations are made, content based recommendations and collaborative recommendations i.e. Similar items are shown or items that people with similar taste had liked would be also be shown. ^[1]

The aim of the report is to find whether there is a relationship between co-purchased items as this would be key to making the recommendation system more accurate which is a common problem with recommendation systems in general.

To achieve this aim, I will be using the “Amazon product co-purchasing network” datasets taken from 12th March 2003 ^[2], 5th May 2003 ^[3] & 1st June 2003 ^[4]. The datasets are provided by the SNAP group at Stanford University. I will also be using SNAP for python ^[5] to analyse the dataset and use gnuplot ^[6] to plot graphs.

I found that buyers of one product will have a narrow range of options from which would be co-purchased as a result. I do feel that I have achieved some interesting analysis and results.

2 BACKGROUND

In this section, I will be describing the key concepts like degree distribution, assortative coefficient & the rich club coefficient that will be used to analyse the datasets.

2.1 DEGREE DISTRIBUTION

The degree distribution is the plot of the frequency of a degree among the nodes in a network versus the degree. For example, for a scale-free network, a power law distribution is shown where there would be fewer high degree nodes than low degree nodes.

2.2 ASSORTATIVE COEFFICIENT ^[7]

The assortative coefficient measures the level of mixing between nodes of similar degree. There are three cases for the assortative coefficient α :

1. α is greater than 0. This implies assortative mixing where similar degree nodes typically link more often.
2. α is less than 0. This shows dis-assortative mixing where higher degree nodes typically link with lower degree nodes.
3. α is equal to 0 which implies the network doesn't show either assortative mixing or dis-assortative mixing.

α is given by the following equation:

$$\alpha = \frac{L^{-1} \sum_i K_i K'_i - \left[\frac{1}{2} L^{-1} \sum_i (K_i + K'_i) \right]^2}{\frac{1}{2} L^{-1} \sum_i (K_i^2 + K_i'^2) - \left[\frac{1}{2} L^{-1} \sum_i (K_i + K'_i) \right]^2}$$

Equation 1 – The assortative coefficient formula

Where L is the number of links, K_i and K'_i are the degrees of the end-nodes of the link I and L is the total number of links in the network.

2.3 RICH CLUB COEFFICIENT ^[7]

The rich club coefficient ϕ , measures how well connected are a group of the richest nodes (richest being the largest degree nodes). ϕ is given by the following equation:

$$\Phi(r) = \frac{2E(r)}{r(r-1)}$$

Equation 2 – The rich club coefficient formula

Where r is the rank of a node based on its degree (for highest degree node, its rank is 1), $E(r)$ is the total number of links between the nodes who rank above or equal to r.

3 LITERATURE SURVEY

In this section, I will be introducing and discussing research works that are related to my work.

3.1 ANALYSIS OF PRODUCT PURCHASE PATTERNS IN A CO-PURCHASE NETWORK ^[8]

This paper aims to study co-purchase networks and derive a recommendation system which would maximize sale revenue. The authors of the paper chose to analyse the same amazon co-purchase networks that I also aim to analyse.

As there is a strong presence of communities in the amazon co-purchasing network, they observed the evolution of these communities over time and found that some communities remained in the network whilst other communities would break and reform over time i.e. reoccurring communities.

The reasoning for the reoccurring communities couldn't be explained as the dynamics of the community that caused nodes to return to a community weren't known.

In conclusion, this paper does provides an interesting idea where communities are one cause for co-purchased items and by knowing these communities, it would improve the recommendation system for Amazon.

3.2 PREDICTION OF CO-PURCHASING PRODUCTS ^[9]

This paper analyses the amazon co-purchasing network and aims to find out the main reasons for co-purchases to happen and then create a model based on their findings to predict future co-purchases.

It differs from the previous paper (Analysis of product purchase patterns in a co-purchase network) in that it uses a different dataset which contains product metadata and product review information.

Their main findings were:

- Popular items (items with more purchasing times) tend to have higher number of co-purchasing items
- High-rated items tend to be co-purchased more often
- Items with earlier first reviews tend to have less number of co-purchasing items
- Similar category items tend to be purchased more frequently
- Similar titled items tend to be purchased more frequently

Following these findings, they constructed a model and found that similar categories and sale rank were the key parameters that enabled better prediction of co-purchase items.

In reviewing this paper, I find that most of their findings were already expected and the key parameters they picked has already been a part of recommendation systems in general and so, isn't that interesting.

3.3 MOTIF ANALYSIS IN THE AMAZON PRODUCT CO-PURCHASING NETWORK ^[10]

This paper aims to observe one of the time snapshot of the amazon co-purchasing network, identify significant recurring patterns of interconnections present in the network i.e. motifs, and describe how these can be used to understand the relationship between products.

The authors identified the following most frequent 3 node motifs:

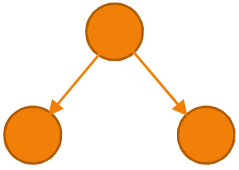
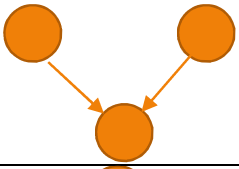
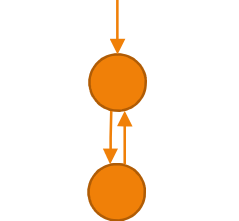
Most frequent		This shows a strongly connected component indicating closeness similarity between the nodes.
2 nd most frequent		This indicates that customers bought unrelated products buys the product below. As there is a high frequency of these, this suggests a majority of customers buy small subset of items.
3 rd most frequent		This is a strongly connected component which displays close correlation between the products where customers buying the product at the top and bottom, may buy the one in the middle.

Table 1: The most frequent 3-node motifs are shown & described in descending order of frequency

The authors then identified the following most frequent 4 node motifs:

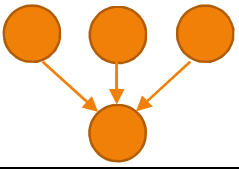
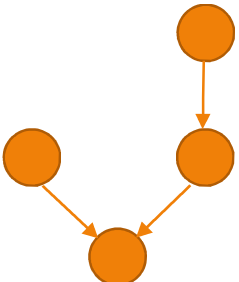
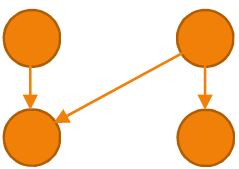
Most frequent		This strongly shows the relationship of convergence to purchasing the below product. Each of these shown the same relationship but as more nodes are in-between, there is a lower probability of purchasing the below products.
2 nd most frequent		
3 rd most frequent		

Table 2: The most frequent 4-node motifs are shown & described in descending order of frequency

In reviewing this paper, it does show some interesting ideas such as how directionality of a link around the neighbourhood of a product can be one method of predicting co-purchases. However, as it was observed during one snapshot in time, it remains questionable to whether the same motifs will continuously appear throughout time.

4 METHODOLOGY

The dataset as described previously are provided by the SNAP group from Stanford University.

I will first observe basic network properties of each of the snapshots in order to see if it can answer my question before proceeding any further.

Following this, I will plot the in-degree distribution and out-degree distribution to find out how the network is structured. I do know that these are plotted in (Basuchowdhuri, Shekhawat and Saha, 2014) which shows a power law relationship however I would like to verify these before proceeding and also, analyse these. If there is a power law relationship as the paper suggests, I will then observe the rich club coefficient and the assortative coefficient to check for the presence of a rich club.

The plot of the in-degree and out-degree distribution can be obtained directly from the python interface provided by SNAP. However, as for obtaining the rich club & assortative coefficient, the python interface provided by SNAP doesn't implement any functions that calculates these. Therefore, I will be programming these and using functions provided by the python interface.

4.1 ASSORTATIVE COEFFICIENT ALGORITHM

For assortative coefficient, the algorithm follows the equation exactly and is as follows:

- Declare and initialise variables to hold the sum of the product, sum of the sum and the sum of the squares
- For each edge in the network,
 - Get degree of the nodes at the ends of the edge
 - Multiply the degrees together and add to variable that holds the sum of the product
 - Sum the degrees together and add to variable that holds the sum of the sum
 - Square each degree, sum them then add to the variable that holds the sum of the squares
- Calculate the numerator of the equation
- Calculate the denominator of the equation
- Finally, divide the numerator by denominator to get the assortative coefficient.

4.2 RICH CLUB COEFFICIENT ALGORITHM

For the rich club coefficient (rank based), the algorithm is as follows:

- Retrieve list of node IDs and node degrees
- Sort the list in descending order of node degrees I.e. node of largest degree is first in list
- Open/Create a file to hold the rich club coefficient dataset

- For all nodes in the list
 - Obtain a graph that includes only the nodes above the rank being observed
 - Get the number of edges
 - Calculate the numerator & denominator of equation 2
 - Calculate the rich club coefficient associated with the rank
 - Calculate the rich club coefficient as a percentage & calculate the rank as a percentage of the largest rank.
 - Write the percentage rank & percentage rich club coefficient into the file
- Close the file

5 RESULTS

5.1 BASIC FEATURES

	Amazon0312	Amazon0505	Amazon0601
Nodes	400,727	410,236	403,394
Edges	3,200,440	3,356,824	3,387,388
Average clustering coefficient	0.4022	0.4064	0.4177
Open Triads	57,851,054 (0.9401)	61,157,198 (0.9393)	60,250,166 (0.9379)
Closed Triads	3,686,467 (0.0599)	3,951,063 (0.0607)	3,986,507 (0.0621)
Diameter	18	20	21

Table 3: Basic network features from three different instances in time. The values in the brackets is the value of that specific triad as a fraction of all triads. ^[2] ^[3] ^[4]

From table 3, it is observed that the average clustering coefficient and the closed triads are increasing with time. As a result of the number of closed triads increasing, the number of open triads are reducing. The diameter of the network is also increasing.

5.2 DEGREE DISTRIBUTION

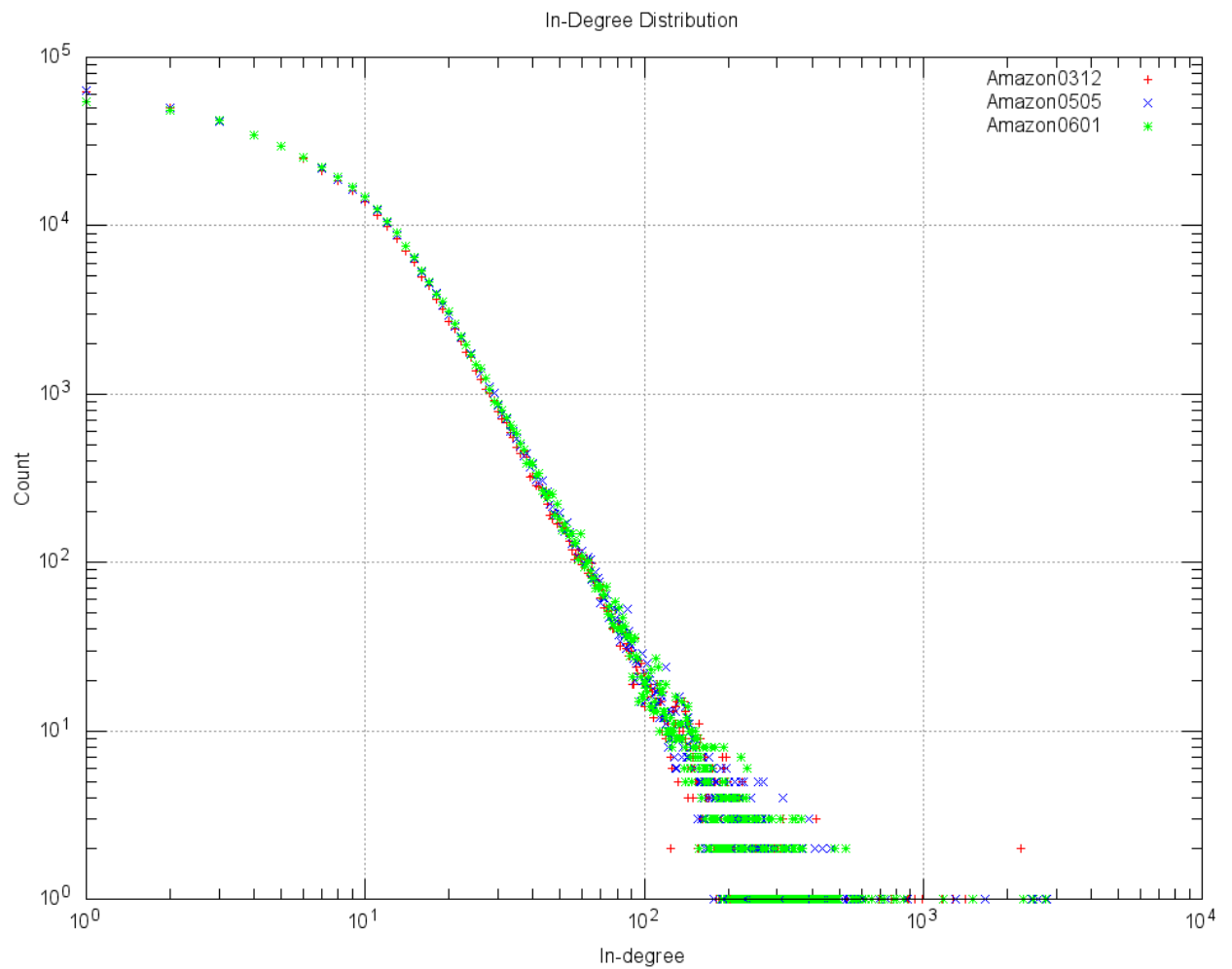


Figure 1: The in-degree of the Amazon co-purchasing network for all three snapshots of the network is shown

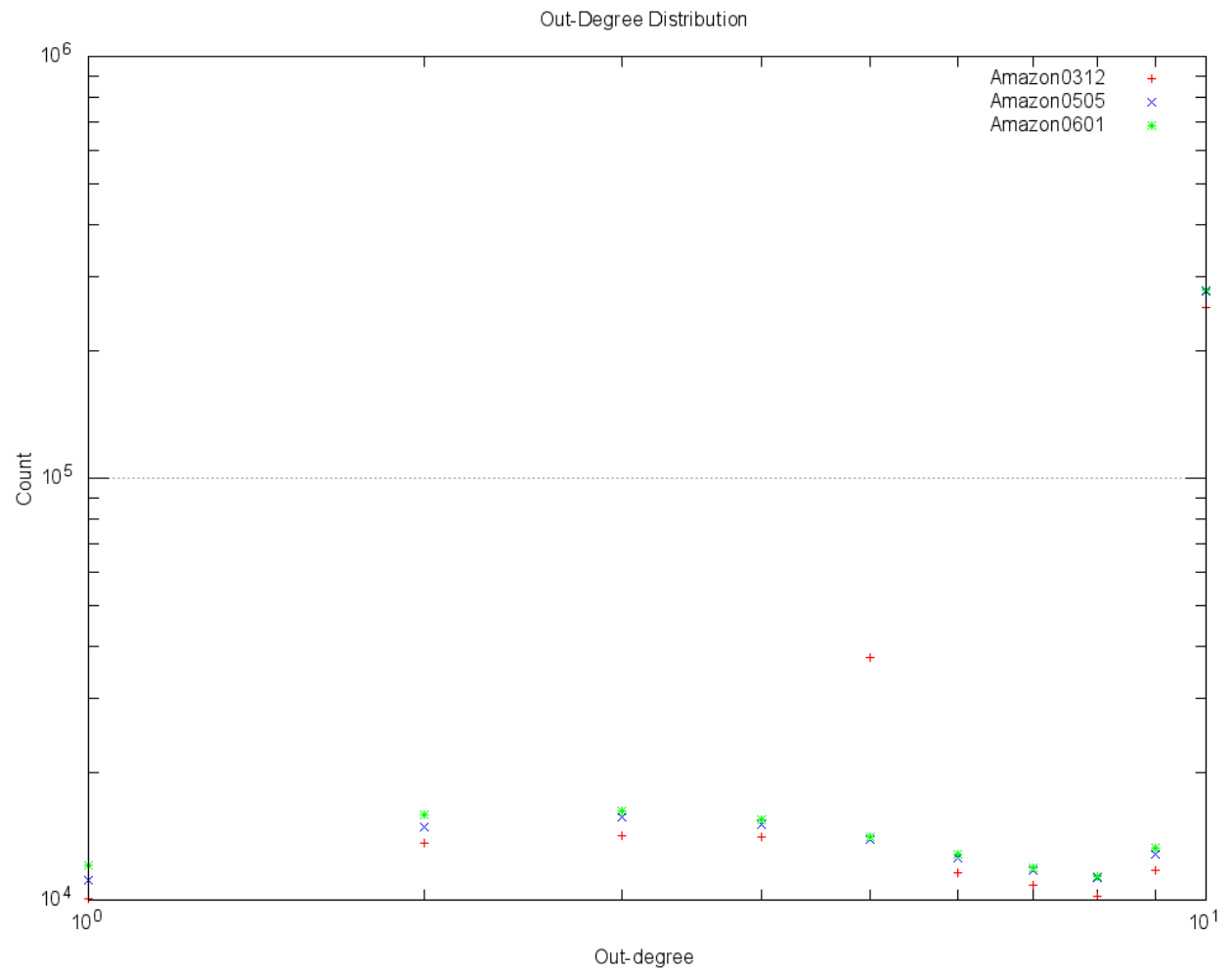


Figure 2: The out-degree distribution of the Amazon co-purchasing network for all three snapshots of the network is shown

As a note, the in-degree is the same as that shown in (Basuchowdhuri, Shekhawat and Saha, 2014) however, the out-degree is not the same which cannot be determined as of why this is the case. In the paper, it shows a power law for the out-degree distribution but for the purpose of this report, I will assume my results are correct.

The in-degree distribution shown in figure 1 suggests the amazon co-purchasing network is a power law network/scale free network and like a scale free network, it has the characteristics of a long tail however, the same cannot be said about the out-degree distribution shown in figure 2.

Another observation is the maximum degree for in-degree nodes is several magnitudes higher than for out-degree nodes.

A final observation which is interesting is that the trend of the in-degree distribution and out-degree distribution remains constant however, the shape of the in-degree distribution remains almost constant across the three snapshots whereas for the out-degree distribution, the distribution becomes slightly larger but maintains the same maximum degree.

5.3 ASSORTATIVE COEFFICIENT & RICH CLUB COEFFICIENT

The assortative coefficient for the amazon network is as follows:

Amazon0312	Amazon0505	Amazon0601
0.298	0.262	0.256

Table 4: The assortative coefficients is shown for the three snapshots of the amazon co-purchase networks

From table 4, the network is seen to be assortative across time but in addition, it can be observed that the assortative coefficients decrease with time.

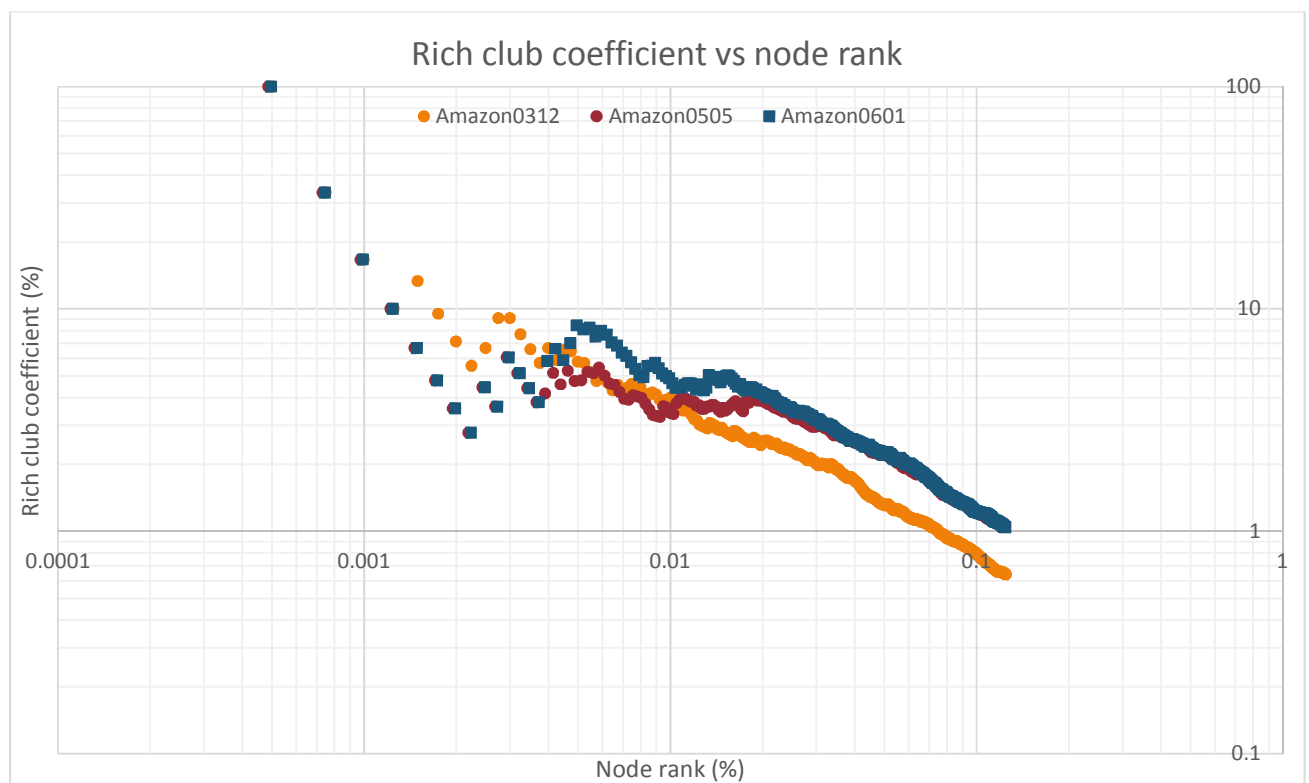


Figure 3: The rich club coefficient is plotted against the degree of a node.

From figure 3, it can be seen that there is no rich club present in the amazon network. For example, observing that only 0.001% of out of the total nodes have a rich club coefficient of 0.1 shows this. The general trend is very similar across all snapshots of the amazon co-purchasing network.

6 DISCUSSION

The fact that both the average clustering coefficient & the closed triads is increasing could show that the network shows evidence for triadic closure however further studies are required to validate this.

Comparing the in-degree distribution to the out-degree distribution, there is a much larger range of degrees (several magnitudes larger) than that of the out-degree distribution. This shows that co-purchased items are common to a range of products that initiate the co-purchasing. As there is a low out degree in general, it also means that people who buy products will be very selective with which products to co-purchase i.e. Most of the time, the choice isn't random.

Additionally, as the maximum degree of the out-degree distribution is the same across all snapshots, this could suggest that there is a maximum of 10 possible products that could be co-purchased with each product.

It is interesting that in a temporal network where products come in and out, trends can disappear and reappear, that the degree distribution remains almost the same whilst the rich club coefficient trend also generally remains the same. Furthermore, as the associative coefficient is decreasing with time, it may be that the network is moving away from being associative in some way but due to lack of data, this can only be assumed.

Returning to the research question, there is a relationship between a product and the products that are bought as a result of it but it appears that there isn't a link between products that have high degrees. On the other hand, it seems as though that there is a finite range of products that are co-purchased which means for future research, it might be interesting to observe what products comprises of these 10 products that are co-purchased with each product.

REFERENCES

- [1] –Tyrchan, C., Falk, N. and Boström, J. (2010). *Exploiting the Amazon.com “People Who Bought ... Also Bought ...” Algorithm in Reagent Selection*. Available at: http://www.eyesopen.com/2010_EuroCUP_presentations/EuroCUP4_Bostrom.pdf [Accessed 15th December 2015]
- [2] –Leskovec, J. and Sosi, R. (2004). *Amazon product co-purchasing network, March 12 2003*. [online] SNAP. Available at: <http://snap.stanford.edu/data/amazon0312.html>. [Accessed 17th December 2015].
- [3] –Leskovec, J. and Sosi, R. (2004). *Amazon product co-purchasing network, May 05 2003*. [online] SNAP. Available at: <http://snap.stanford.edu/data/amazon0505.html>. [Accessed 17th December 2015].
- [4] –Leskovec, J. and Sosi, R. (2004). *Amazon product co-purchasing network, June 01 2003*. [online] SNAP. Available at: <http://snap.stanford.edu/data/amazon0601.html>. [Accessed 17th December 2015].

- [5] – Leskovec, J. and Sosi, R. (2004). *SNAP for Python* (Version 1.2) [Computer program]. Available at: <http://snap.stanford.edu/snappy/index.html> [Accessed 23rd December 2015].
- [6] – Williams, T. and Kelley, C. *Gnuplot* (Version 5.0) [Computer program]. Available at: <http://www.gnuplot.info/> [Accessed 26th December 2015].
- [7] – Zhou, S. (2015). *Mixing Pattern and Rich-Club*.
- [8] – Basuchowdhuri, P., Shekhawat, M. and Saha, S. (2014). *Analysis of Product Purchase Patterns in a Co-purchase Network*. Available at: <http://doi.ieeecomputersociety.org/10.1109/EAIT.2014.11> [Accessed 20th December 2015].
- [9] – Liu, Y., Wu, C. and Tong, X. *Prediction of Co-purchasing Products*. Available at: <http://cseweb.ucsd.edu/~jmcauley/cse190/reports/039.pdf> [Accessed 20th December 2015].
- [10] – Srivastava, A. *Motif Analysis in the Amazon Product Co-Purchasing Network*. (2010). Available at: <http://arxiv.org/abs/1012.4050> [Accessed 20th December 2015]