Instituto Superior Técnico

Network Science

2021/2022 – 1st Semester (MEIC-A, MEIC-T, MECD, MMA)

Recommendation Network – Amazon

Group 44

Ricardo Ferreira 87701, João Jorge 88079

A person who bought product ‘x’ also bought product ‘y’. Knowing this, a recommendation network was created by Amazon compiling consumer’s purchases and arranging them in a determined way. In this brief paper we discuss this network, our analysis of it, and present various graphs that quantitatively illustrate our findings. We looked at some of the network’s properties, particularly as they relate to most bought products as well as communities where products belong. Our findings may help Amazon predict how future customers will behave. We believe that this type of recommendation network study and analysis can also help Amazon in planning the quantity of its inventory as well as the internal organization of its warehouses.

Introduction

Imagine entering a store and the products on the shelves are arranged so that ones you like or need appear immediately in front of you. Then imagine that for each person in the world there is one store arranged personally for them. The goal of a good recommendation network is to achieve this personalized result, and in that way, enhance the profit of the company. As Amazon stated “We aim to be Earth’s most customer centric company. Our mission is to continually raise the bar of the customer experience by using the internet and technology to help consumers find, discover and buy anything (…)”[1].

Knowing this, a study of the recommendation network topology can help the company direct its efforts in smooth and fast delivery of a new requested product by storing possible products and arranging them in a proper way. This study can also focus Amazon’s marketing efforts on most bought products and their respective communities.

We divided the work into four parts, always completing a general analysis of the netwo rk topology and then relating those measures with the products in the network.

[1] By Amazon Staff. *Our Mission.* https://www.aboutamazon.co.uk/uk-investment/our-mission

We first discovered that the recommendation network is for all intents and purposes, pra ctically connected (403,364 products connected compared to 30 products disconnected). This mean it is possible to reach any part of the network if one product is bought, makin g it reliably possible to follow a particular chain of products. This demonstrates a certain amount of network resilience and suggests that Amazon does not need to link products m anually or unite clusters and categories to offer a wide variety of different products. We t hen calculated the average shortest path between products and found that on average, a p roduct is 6.5 hops away from any other product, which again supports the claim made ea rlier.

This network seemed well organized although no human has programmed it to behave li ke this. To provide a more quantitative view on this point, we calculated the degree distr ibution and found this measure to show what we suspected. The graph has some propert ies of a scale free network. We say some properties because the most bought product is n ot as bought as it would be in a typical scale free model. The most bought product has 2,752 purchases while it would be expected to have around 100,000 if the network perfe ctly fit the mathematical model. We note that 20% of the most purchased products accou nt for 42% of total sales. We do not know if the network is stable or if it is still evolving to a more perfect scale free network, since we did not compare this network with past or future models.

To complete our study, we also ran a community finding algorithm and tried to figure out how products aggregate. The products seemed to arrange randomly between subjects, which we were not anticipating. Instead, they aggregate on the number of sales.

Results

The recommendation network determines how a virtual store is organized and what products one is probably going to buy [2].

We decided to start as broad as the statement seemed to us, using a very general overview of the graph, and it was a very profitable way of doing so! Almost all nodes are connected, leaving out only 30 of the total 403,394 products. These 30 disconnected products group into 6 different communities, each having 15, 5, 3, 3, 2, 2 products. We looked at each small community and at this time were unable to determine why they are separated from the larger cluster. It may be because they had been recently added to the network and no links had yet formed. Though we think the most probable cause is because the network is in the *supercritical regime* and very close to the *connected regime*.

One of the communities of 3 elements has all products of the same category (Book) and of the same subject (Religion and Spirituality) but the other communities appear to have nothing in common. For example, in the community of 15 elements we have music with the subject ‘Hard Rock and Metal’, a book of science and mathematics, a children’s book of religion and spirituality, a classic music book, and a book with the title *'10,000 Ways*

[2] Smith, B., & Linden, G. (2017). Two decades of recommender systems at Amazon. com. *Ieee internet computing*, *21*(3), 12-18.

to Say I Love You: The Biggest Collection of Romantic Ideas Ever Gathered in One Place' (For more information about the products of each community see appendix [1]).

As we noted, we suspect these products will join the main component if the network follows the model of a random graph or scale free network. Regardless of not finding order in the communities outside the main component, we conclude some order is present in the network because the products connect to each other without external interference.

Returning now to the main component, it is interesting to note that this characteristic of the network tells us that it is possible to reach any type of product by following a certain chain of purchase. Following this reasoning the next lead we pursued was finding the shortest average path. We got the result of 6.5 hops! With this measure, we know the average distance from one product to another in the network. We also compared the shortest distance between two randomly different categories (for example: Book, Music), different subjects (Computer Science book, Sports book, Religion & Spirituality book, Health, Mind & Body book) and same subjects (like sports). The shortest distance found was a bit higher than average for books in different categories and different subjects (6.8 and 7.2 respectively) and smaller than average for products with the same subject (5.4 hops on average). We have to point out that these distances were calculated with a small sample (5 products for each measure) and so the reality of the network may differ a bit from these results (see appendix [2]). That is why we took no conclusions from these facts, even though they are interesting and show promising results if pursued at length.

We also looked at how the shortest distance behaves between nodes with high degree distribution, low degree distribution and these two metrics combined. The results were very interesting: between nodes with high distribution the average shortest distance was 2.2 hops and for the nodes with low degree distribution it was 7.8. Then we calculated the distance between a node with high distribution and one with low. The average distance was 5.3 hops, which is smaller than the average 6.5 (*Figure 1*).

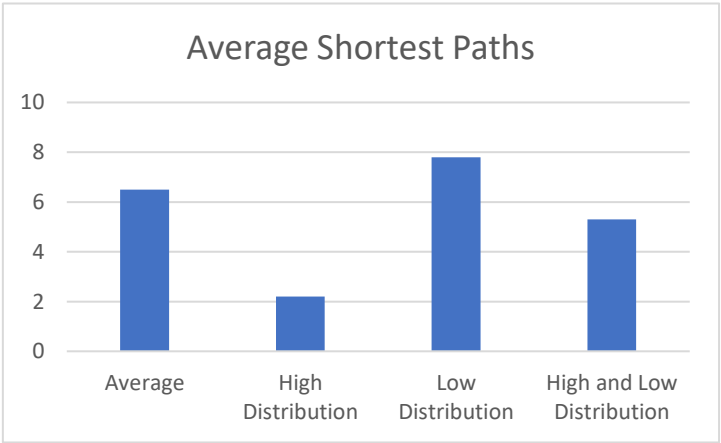

Fig. 1 - Difference between average shortest distance in products with different degree distribution

We concluded this metric noting the following: products that are purchased repeatedly, have a very short distance between others with high distribution, and are closer on average than all others, even products with low distribution. Products rarely purchased are further away from other low purchased products, but the distance is not that different from the average once they have easy access to highly bought products, and from there every other product can be reached.

With these two measures, the existence of only one community and the average short distance, we can see that the recommendation network organises itself in a very determined way. After we got these results the idea of a scale free network started to appeal to us as a probable reality, so we plotted the degree distribution of the graph and got an interesting result (see figure 2).

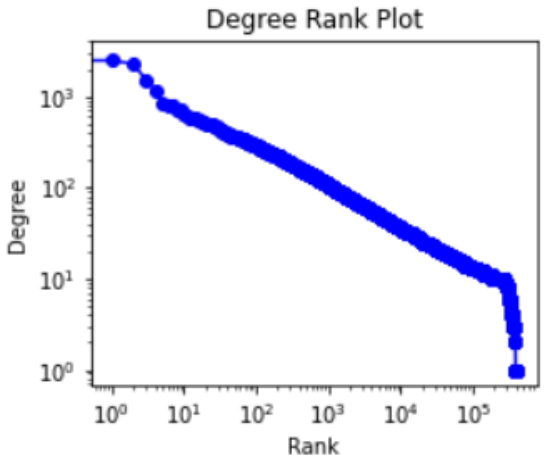

Figure 2 - Degree distribution of the network. 'y' axis has number of links and 'x' axis number of nodes (log-log scale)

It appears to be a scale free network, but not a very clear one. Instead of the normal 80/20 partition, which means 20% of the products are responsible for 80% of the sales we got that 20% of the products (80,678 in cardinal) are responsible for 42% of the total sales. This is perceptible in the graph if we look at the ‘x’ and ‘y’ axis (note that the number of purchases is represented on the ‘y’ axis and the node number on the ‘x’ axis). The product with most sales has 2,752 purchases and it quickly decays to 200 sales per product. In fact, the 1,001 most bought product has only 107 purchases (this product is in the 0.3% of the most bought). It is also very interesting to note the right part of the plot shows a rapid decay of sales. The least bought product has only 1 [one] sale and more than 11,000 products belong to that sole point (11,000 products correspond of 2.7% of total products). These last findings actually impact significantly as to how Amazon looks at the chain of purchases.

Because this is not a perfect scale free network, there is a considerable number of products that are widely bought (not only the highest bought products) and it is not difficult to introduce new products to the market at any time since there is no true market monopoly for any product. For a product to be considered successful, it has to pass the low barrier of 10 sales.

It would be interesting to look at how time affects the network, but we decided to move on to analyse the communities of the network and discover how and why products aggregate together. But if one is interested in following 'the timing of recommendations and purchases' it might be interesting to look at Chapter 7 of "The dynamics of viral marketing"[3].

To complete our study of the network we looked at how products link up together. We used the community finding Louvain algorithm[4] to aggregate products into communities. It appears the subject of the product is not connected with the community where it belongs (using the referenced algorithm). For example, in one community with 68 products, there were 40 different categories (figure below).

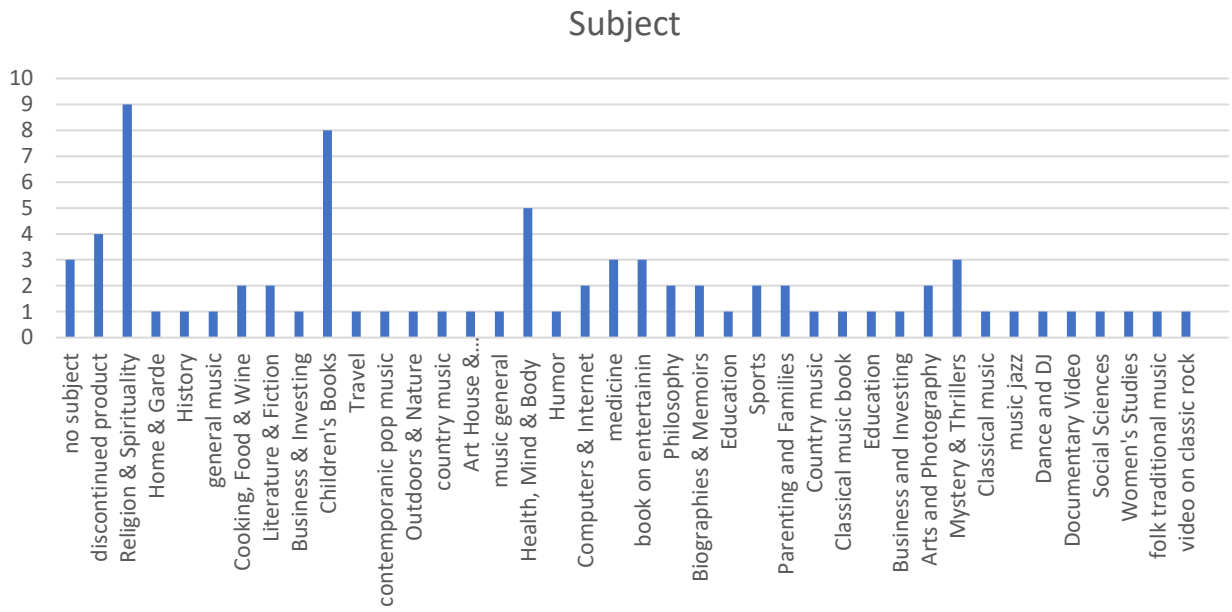

Fig. 3 - Subject of products in products within the same community

Even for products with the same subject the title often suggests that there is no relation between them. Note for example, two books belonging to the same community and with the subject, *Religion and Spirituality*. One book has the title 'Patterns of Preaching: A Sermon Sampler' and other 'Dolphin Talk: An Animal Communicator Shares Her Connection'. This is illustrated further with another example in a community with 4 elements having the subjects: *Physics Sciences* (title: 'Theory of sound'), *a rock music* (title: 'At the edge'), a book on *science fiction and fantasy* (title: 'Hokas Pokas!') and a book with no subject registered (title: 'The Real James Herriot : The Authorized Biography').
We continued to look further at other communities and discovered no relation of subjects between products. We then pursued the path of registering degree of nodes within a

[3] Leskovec, J., Adamic, L. A., & Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Transactions on the Web (TWEB), 1*(1), 5-es.
[4] Community detection for Networkx. https://python-louvain.readthedocs.io/en/latest/api.html

community. The results were much more promising and we concluded that the clusters are formed in communities with the same degree and not because of subject nor category similarity. For a community with 32 elements the lowest product degree is 2 and the highest is 28 which is not a big diference since there are thousants of products with more than 300 sales and difference between the highest degree node and the lowest is only 26 purchases (figure below).

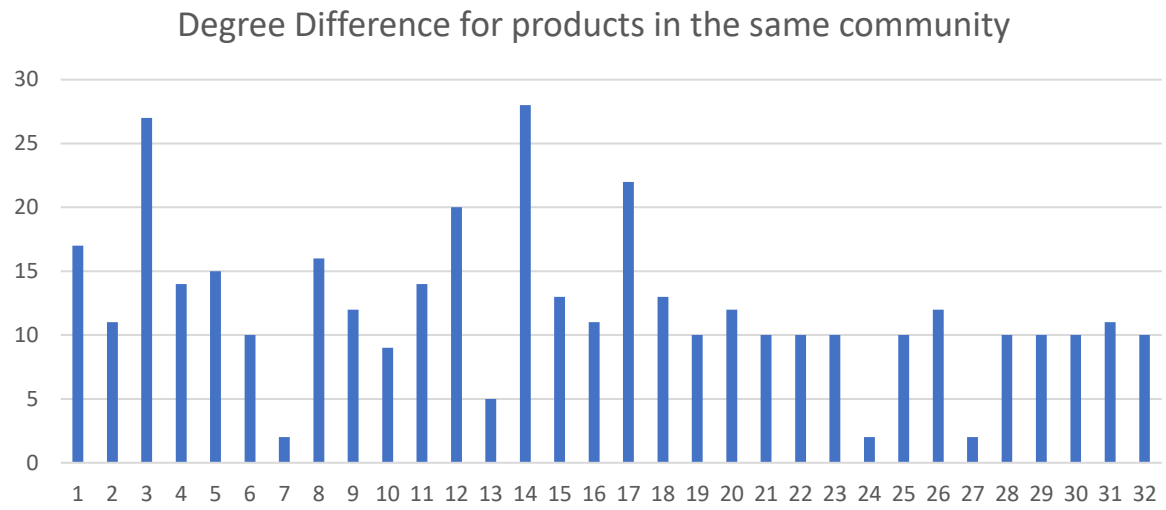

Fig. 4 – Degree difference for products in the same community. Note that there are lots of products with 10 sales. The other do not vary much more. Vertical line is the number of sales (degree) and horizontal line, the product in the community.

For 403,394 products the community finding algorithm returned 3 levels of partitions: the first one having 28,425 communities, the second 2,087 and the third 256.

Conclusion

Targeting some specific products for marketing or warehouse organization is not as simple as we initially thought, since the network presented is not a perfect scale free model. Nevertheless, we were able to obtain a number of interesting insights on the average distance of purchase between products and the strong main component that the network is built around. The products' aggregation did not return a pattern, according to category or subject as intuition dictates but rather to the number of sales. It turns out that popular products unite with other popular products.

Our analysis suggests that potential studies can be done, especially those focused on the evolution of purchases in time. It also might be worthwhile to try other methods of community finding and see if there are other measures to consider alongside with the number of sales. We say this because some products in communities appeared to have a very close identification number which indicates similar purchase time or even same order purchase time (see appendix [3]).

Methods

It is possible to build the recommendation network in two different ways: similarity between customers, meaning its preferences and likes; or correlation between products, meaning who bought product 'x' also bought product 'y'. Amazon found that analysing correlation between products produces a better recommendation network[5].

We then got the recommendation network from the dataset collection[6] with the description "Amazon product co-purchasing network from June 1, 2003". It is a simple ".txt" file where there are 403,394 nodes and 3,387,388 edges. The products' identification (title, subject) is in a separated file on the same website. Products are divided into the following categories: books, DVDs, music, and video.

The file contains the links between nodes. It is a directed graph because if product 'x' is bought with product 'y' it does not mean that product 'y' is bought with product 'x'. We decided to work with the undirected graph because it would facilitate some graph operations and diminish the complexity of some algorithms. We ended up with 2,443,408 edges instead of the initial 3,387,388. We wrote the code in python and mainly used the *networkx* (4) library. The code is referenced so that it is possible to reach supporting documentation with ease. Note that only for the shortest average path, the function that calculates the average shortest distance between all nodes was not used due to time constraints (it would take several days to compute that measure). Instead, we chose 200,000 random nodes in the network and computed their shortest distance.

For the degree distribution, the plotted graph looks like figure 4 and in log-log scale looks like figure 5 (the same as figure 2 in page 3).

We used the community Louvain for community finding and printed the results in ".txt" files.

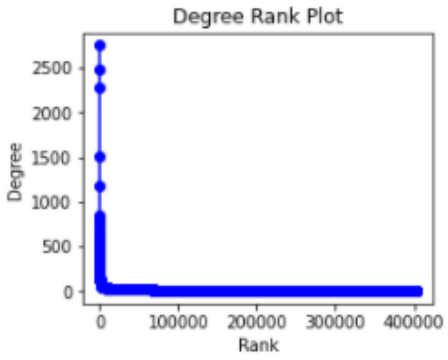

Figure 4 - Degree distribution of the network. 'y' axis has number of links and 'x' axis number of nodes.

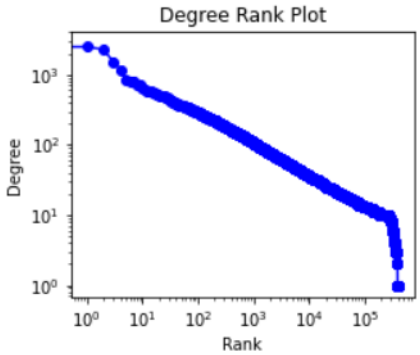

Figure 5 - Degree distribution of the network. 'y' axis has number of links and 'x' axis number of nodes. Note that it is a log-log scale.

[5] Hardesty, L. (2019, November 22) The history of Amazon's recommendation algorithm. https://www.amazon.science/the-history-of-amazons-recommendation-algorithm

[6] Standford Large Network Dataset Collection. https://snap.stanford.edu/data/index.html#amazon

Appendix

```
#2 elements group{340003, 340004}
#music, Rock, 'Living Off the Radar'), (book, Literature and fiction, 'My Antonia')

#2 elements group {360296, 360297}
#(video, Actors and Actresses, 'The Monkees - Our Favorite Episodes (in Metal Lunchbox)')
#(Book, no categories, 'Beatrix Potter Mom's Brag Book')

#3 elements group {376281, 376282, 394995}
#(Book, Cooking, Food and Wine, '100 Innovative Recipes--From Appetizers to Desserts')
#(Book, no categories, 'The Maltese Dog (Wishbone Mysteries Promotion , No 6)')
#(Music, Classical, 'Sofia Gubaidulina: Offertorium (Concerto for Violin & Orchestra, 1980) /
    #Hommage à T.S. Eliot, for Octet & Soprano (1987) - Gidon Kremer / Charles Dutoit')

#3 elements group {383840, 383841, 383842}
#(Book, Religion and Spirituality, 'The Tabernacle of David')
#(Book, Religion and Spirituality, 'God's Life in Us')
#(Book, Religion and Spirituality | Philosophy, 'The Inner Reaches of Outer Space: Metaphor as Myth and as Religion')

#5 elements group {385301, 385302, 385303, 385304, 385305}
#(Book, Literature and Fiction | Romance, 'Real for Me')
#(Book, Literature and Fiction, 'Oh My Goth!')
#(Book, Crafts and Hobbies | Home and Garden, 'The Art of the Quilt')
#(Book, Business and Investment, 'The Engaged Customer: Using the New Rules of Internet Direct
#    Marketing to Create Profitable Customer Relationships')
#(Book, Sports | Basketball, 'Three-Point Field Goal Offense for Mens and Womens Basketball
#    (The Art and Science of Coaching Series)')

#15 elements group {384064, 384065, 384066, 384067, 384068, 384069, 384070, 384071, 384072, 384073, 389697, 394986,
#    394985, 386528, 384063}
#(Book, Children's book | Religion | Christianity, 'The Miracles of Jesus (Young Reader's Christian Library)')
#(Book, Home & Garden, 'Ruffing It: The Complete Guide to Camping With Dogs')
#(Music, Classical | Opera,'Japanese Melodies')
#(Book, History | Military, 'The American Civil War: The War in the East 1863-1865 (Essential Histories)')
#(Book, Literature and Fiction | Erotical | Adult Fiction'The Lov-Ed Solution')
#(Music, Dance and DJ, 'Mixed')
#(Music, Classical, 'Michael Tilson Thomas Performs and Conducts Gershwin')
#(Music, Classical | Opera and Vocal,'Fascinatin' Rampal (Jean-Pierre Rampal Plays Gershwin)')
#(Music, Latin Music, 'Rabanes')
#(Music, Featured Composers, 'Haydn: Flute Concerto; Oboe Concerto')
#(Book, Science | Mathematics, 'Foundations of Mathematical Logic')
#(Book, Home & Garden, 'R 2800: Pratt & Whitney's Dependable Masterpiece [R-241]')
#(Book, Social Sciences, 'Intercultural Competence: Interpersonal Communication Across Cultures (4th Edition)')
#(Book, Health, Mind and Body, '10,000 Ways to Say I Love You: The Biggest Collection of Romantic Ideas
#    Ever Gathered in One Place')
#(Music,Hard Rock and Metal, 'Real Life')
```

Appendix 1 – communities outside of the main component. Here we present the 6 communities, each having the respective id, category (book, music, dvd), subject (History, Religion) and title.

```
#Compare different categories
#different categories (book, 66244) and (Music, 327220) = 8
#different categories (book, 171441) and (Music, 327220) = 8
#different categories (Music, 160760) and (DVD, 253209) = 6
#different categories (DVD, 160729) and (Book, 96365) = 6
#different categories (Video, 272471) and (Book, 322875) = 6
#average = 6.8


#compare different subjects
#different subjects (Computer Science, 99093, Book) and (Cooking, Food & Wine, 206462, Book) = 8
#different subjects (Sports, 393195, Book) and (Computer Science, 206466, Book) = 7
#different subjects (Cooking, Food & Wine, 393327, Book) and (Health, Mind & Body, 149251, Book) =8
#different subjects (Health, Mind & Body, 149244, Book) and (Sports, 216862, Book) = 6
#different subjects (Biographies & Memoirs, 376670, Book) and (Religion & Spirituality, 1, Book) = 7
#average = 7.2


#compare same subjects
#same subjects (Computer Science, 99093, Book) and (Computer Science, 206466, Book) = 6
#same subjects (Cooking, Food & Wine, 206462, Book) and (Cooking, Food & Wine, 393418, Book) = 7
#same subjects (Sports, 216862, Book) and (Sports, 393412, Book) = 6
#same subjects (Health, Mind & Body, 149251, Book) and (Health, Mind & Body, 149244, Book) = 1
#same subjects (Biographies & Memoirs, 393391, Book) and (Biographies & Memoirs, 334866, Book) = 7
#average = 5.4
#average = 6.5 (discarding value 1)
```

Appendix 2 – average short distance between subjects. Divided in three sections, we compared the average distance between products with different categories, different subjects and same subjects. The results were not conclusive because the sample of products is small.

```
#community 6952 (2 elm) : [105166, 185103] = [5,5]
#community 28407 (2 elm) : [395878, 399522] = [2,1]
#community 21047 (2 elm) : [402702, 402703] = [2,2]
#community 20387(5 elm) : [321943, 366092, 366093, 366094, 366095] = [10, 4, 5, 4, 4]
#community 28203 (9 elm):[68302, 68303, 68304, 68305, 73508, 130466, 130467, 179445, 258632]=[18, 14, 16, 15, 11, 10, 11, 10, 1]
#community 24498 (32 elm) : [63631, 51821, 68586, 54911, 54912, 54913, 54915, 63784, 63789, 63790, 74407, 68585, 74405, 115669,
#115674, 259854, 115668, 115670, 115671, 115673, 115676, 184568, 160547, 160548, 160549, 176786, 299555, 176784, 177568, 298363,
#239346, 344204]
#= [17, 11, 27, 14, 15, 10, 2, 16, 12, 9, 14, 20, 5, 28, 13, 11, 22, 13, 10, 12, 10, 10, 10, 2, 10, 12, 2, 10, 10, 10, 11, 10]
```

Appendix 3 – degree distribution within products in the same community. Th community '6952' has 2 elements: the elements with identification '105166' and '185103'. Their degree is [5,5] respectively.