Analysing the Amazon Co-purchasing Network

Noshad Boksh
210600322
Network analysis methodology

Jun Hann Chong
210586749
Results and discussion

Remy Philip Jean MacDermott
210334586
Dataset and network presentation

Senthuran Krishnakumar
210344862
Related work

Abstract—**This report is an investigation of the Amazon co-purchasing network, wherein Gephi is used to analyse the complex network it features. By using a cleaned dataset, the necessary small-world network properties are identified with ease and accuracy. Central hubs and product communities may lead to network vulnerabilities, by which means of processing the cleaned dataset attempts to abolish this potential breach. The research conducted for this report uses a set of network analysis methods and techniques, such as performing a PageRank analysis on the data to gain an understanding of Amazon's marketing strategies via their complex co-purchasing network.**

I. INTRODUCTION

The field of digital media and social networks continue to grow at a rapid rate, bringing new innovations and findings at an admirable pace. This report attempts to contribute to this expanding field by navigating the complexity of the Amazon co-purchasing network, utilizing an extensive dataset. In brief, co-purchasing can be described as the system of recommending related products in reference to products in a basket, recently purchased/viewed products or the product on the current webpage. The issue we face can be described as a challenge in comprehending intricate systems. Through analysing the current dataset, we aim to gradually broaden our foundational understanding of these complexities. As such, the challenge to address is to use a cleaned dataset of Amazon products to gain an understanding on how such co-purchasing system works and the degree of accuracy which exists in it.

Using Gephi, a powerful network visualization tool, the network structure of the co-purchasing is analysed in great depth to pinpoint any existing patterns. By identifying noticeable clusters in the graphs, conclusions can be made into helping outline a set of characteristics and points that will help in understanding the marketing excellence in Amazon's system. The analysed dataset includes an extensive range of data from March of 2003 which was cleaned for enhanced results- ensuring that any unwanted data or perhaps unrelated data that would not be in close accuracy when analysing the rest of the data is removed.

In completing this, we concluded that the amazon co-purchasing network exhibits small world network properties that enables efficient navigation and specific recommendations because of the short path lengths yet high clustering. However, potential vulnerabilities are also created as network resilience and echo chamber effects since there is an over reliance on central hubs and the product communities being tightly-knit.

.

II. RELATED WORK

In their paper, Liu, Wu and Tong [6] have conducted an extensive analysis of a similar predictive system. The relevance and need for co-purchasing products are outlined in the field of online shopping as it greatly enhances revenue due to higher sales, furthermore, using how category similarity is a major factor in this and many similar recommending systems. A similar dataset was used by them. Basic characteristics and product groups were identified along with pair analysis, which was expanded upon via page rank analyses. This enabled a correlation to be formed as to how the co-purchasing system works. Data such as sales rank and item rating are some of the metadata which was used to find any relevant similarities between products and the way the co-purchasing system recommended related products. Insights from the analysis are finally provided, which concludes the findings of how category similarity plays a huge role in the recommending system (by means of k-nearest neighbours algorithm).

Prasad, Kumari, Ganguly and Mukherjee [7] in their study discovered that the traditional methods of co-purchase predicting are highly based on product review systems in place. A much better method is using co-purchasing networks as discussed in this report. In their study, the amazon products are nodes, where links signify frequently bought together items. Each item's centrality corresponds to the sales-rank and as a result the system works without requiring reviews. Furthermore, newly added products will therefore not be affected by the lack of reviews and user interaction and can still be used in recommending other products. They also found that the CD and phone results had a greater degree of accuracy when the traditional method was absolved. This can perhaps be attributed to the fact that the metadata of such product is easier to compare with others.

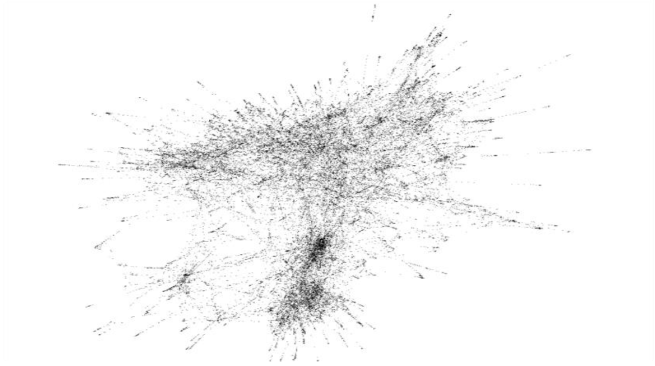
Basuchowdhuri, Shekhaway and Saha [8] studied a similar dataset for the Amazon co-purchasing network. Their findings can be summarised as having chosen central entities within the network which was used as the foundation to suggest similar products. This was done by analysing nodes with high in-degree and out-degree importance. By singling out clusters in the resulting graphs and findings, a frequency degree was found in the co-purchasing system. Using the collected nodes, in other words cleaned dataset, the “evolution of communities in the network” could be analysed by pinpointing changes in association. This was a study conducted in order to find how co-purchasing systems and their accuracy could benefit the company by increasing sales of certain items.

III. Dataset and network presentation

TABLE I. TABLE TYPE STYLES

Node ID	Degree	Out-Degree	In-Degree
481	91	5	86
99	74	5	69
18	68	5	63
8	68	5	63
33	68	5	61

Our report looks at the amazon co-purchasing network of March 2nd 2003. It allows us to visualise and better comprehend the purchasing trends on amazon at the time. It does this by having each item bought be represented as a node with directed edges representing items that a node was also bought with. The analysed dataset includes an extensive range of data from March 2003, which was cleaned to enhance results, ensuring the removal of any unwanted or unrelated data that might not closely align with the rest of the dataset during analysis.



Usings Gephi's powerful tools we are able to analyse many parts of the network to further our understanding of its underlying structure. Gephi also helps us visualise the network using numerous different algorithms to organise the nodes we see. Above we have the network when put under a Force Atlas 2 layout algorithm. The main principle for this algorithm is that nodes connected by edges will attract each other and move away from nodes which aren't connected by and edge. This allows us to more easily visualise communities and clusters withing the network. In the above graph we can clearly see where nodes group together forming communities. The areas that look denser are the areas where items are often bought together for example and Knife and Fork or other items that would commonly be associated with another. In the less dense areas, we have items that aren't frequently bought together. Our layout facilitates clear visualization of dense communities, enabling us to identify clusters where recommending related items would be advantageous. Gephi also helps us analyse other aspects of the network such as the degree of each node which is the number of incoming and outgoing edges on a node. We can obtain the top 5 nodes in terms of degree.

Node ID	Degree	Out-Degree	In-Degree
481	91	5	86
99	74	5	69
18	68	5	63
8	68	5	63
33	68	5	61

The out-degree is the number of edges pointing out of the node and the in-degree is the number of edges pointing to the node. Here we can see how node 481 is purchased with many other items. This suggests that we should prioritise recommending this item when another of the 86 items that are bought with it are purchased. This is because its high degree means its frequently bought together with other items which puts forward the idea that this specific item is ideal for co-purchasing.

We can also look at other values such as the closeness centrality. The closeness centrality allows us to measure the importance of a node relative to its proximity to other nodes. We can also have a look at the betweenness centrality which is a measure of the extent to which a node acts as an intermediary to other nodes. More specifically betweenness centrality measures the number of shortest paths between pairs of nodes in the network that pass through a particular node. Nodes with high betweenness centrality are those that lie on many of these shortest paths, acting as critical connectors between various parts of the network. Below we can see the tables for the 5 nodes with the highest closeness centrality and betweenness centrality.

Node ID	Closeness Centrality
9383	1
8467	1
7266	1
5994	1
3317	1

Node ID	Betweenness Centrality
117	14332021
97	13200516
18	10837374
132	10273133
8	10205722

As we can see multiple nodes have a closeness centrality of 1 which indicates they all have equally short average path lengths.

IV. Network Analysis Methodology

Analysis of Network Properties

The project aimed to explain the underlying structure of the Amazon co-purchasing network by hypothesizing the presence of either small-world features which indicate efficient information distribution and connectivity, or random network features. To gain a better understanding of the typology of the network and its implications for the efficiency and effectiveness of the recommendation system, our analysis focuses on compared these observed network metrics with standard models.

Methodology and Comparative Analysis

Using Gephi, a visualisation and exploration software for all types of graphs and networks, we calculated and visualised the average path length and clustering coefficient, key indicators of small-world networks. We were able to determine the typology of the Amazon co-purchasing network by comparing these measurements to the values of an expected random network of similar size, using the Erdős-Rényi model as a benchmark.

This foundational analysis sheds light on how customers interact with and browse Amazon products, offering important insights into the strength and effectiveness of the network's topology. Determining if the network has random or small-world properties is crucial as it directly influences the flow of information through the network, impacting product discovery and overall user experience.

PageRank Analysis with NetworkX

We applied the PageRank algorithm using the NetworkX library to assign importance scores to products. Through the identification of products that are essential to the networks structure and therefore likely to have an impact on consumer purchasing patterns, this method measures the influence of individual products based on their connections within the co-purchasing network.

Justification for Methodological Choices

The strategic deployment of the PageRank algorithm for identifying influential products in the network is supported by its demonstrated versatility and effectiveness across various network types. [4] has adapted the PageRank algorithm to assess node relevance in directed-weighted networks validating its application in our analysis and highlighting its broader applicability beyond web search engines. In addition, [5] application of PageRank to an air traffic control network, where its performance outperformed alternative ranking techniques, validates its choice for identifying important nodes in the Amazon co-purchasing network. These adaptations demonstrate PageRank's robustness when examining complex structures, which makes it a perfect tool for our investigation into improving Amazon's product suggestion algorithms.

Community Structure Analysis

We then looked at the community structure in the Amazon co-purchasing network as part of our research. The goal was to find and comprehend the product "communities" or clusters that consumers regularly buy together. This knowledge is essential for understanding market niches and improving focused recommendation systems.

Community Detection Methodology

We divided the network into communities for this analysis using Gephi's modularity tool, which applies the Louvain technique for community detection. This method finds divisions in the network that maximise the density of intra-community edges and minimise the density of inter-community edges by optimising the modularity score. Gephi provided a visual interface to explore these communities, highlighting patterns of co-purchasing behaviour among groups of products.

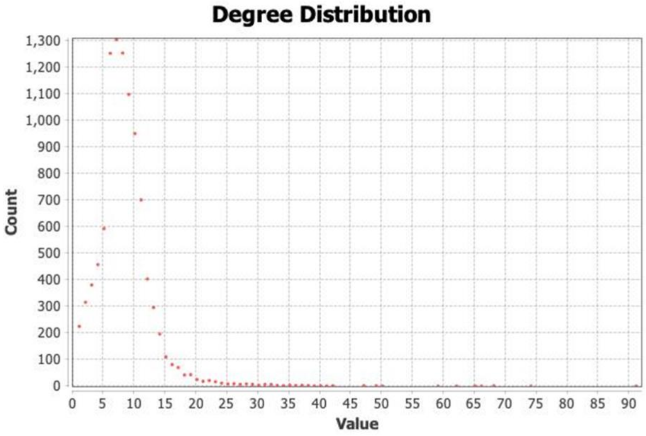
Visualisation and interpretation

We successfully managed to display the network's community structure in a clear and understandable way by using Gephi's visualisation features. Through these visualisations we could observe the size and interconnectedness of communities, allowing us to carry out a qualitative assessment of the network’s segmentation.

Justification for Methodological Choices

The decision to use the community detection algorithms is reinforced by their demonstrated ability to refine and enhance e-commerce recommendation systems. The thorough analysis by [3] emphasises the use of community detection in hybrid recommender systems to overcome the shortcomings of traditional recommendation methods. Additionally, [2] exploration into the formation of user communities through community detection aligns with our approach, supporting the hypothesis that such techniques can substantially improve the diversity and precision of product recommendations. These community detection applications across e-commerce contexts support our methodological approach and demonstrate how the knowledge gained from this analysis can greatly improve Amazon's ability to offer personalised product recommendations, improving the overall shopping experience for consumers.

V. Results and discussion

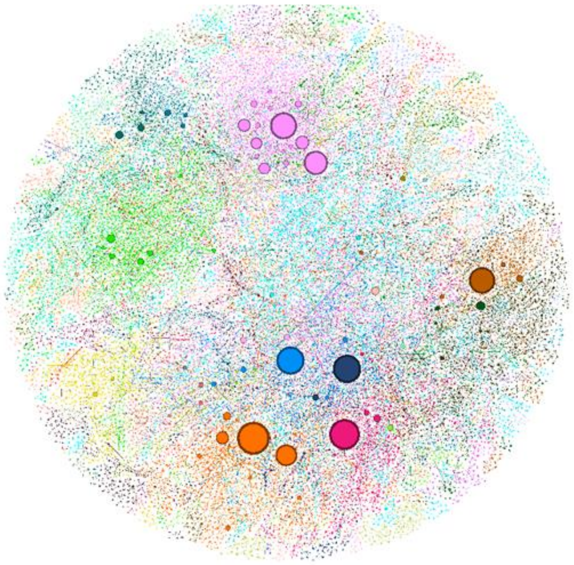


In calculating the degree distribution, it allows us to view insights into the structure and dynamics of network. In context of the Amazon product co-purchasing network, understanding degree distribution can inform us on how products are linked based on customer purchasing behaviour. From the results of our degree distribution calculations, they reveal a heterogeneous network, usually distinguished by a heavy-tail distribution as seen in the graph above. This characteristic indicates the presence of few large hubs with many products linked to them, with majority of products having few connections. This pattern can typically be found in scale-free networks, where many products are often co-purchased with a few top selling products. Networks with the presence of hubs also indicate strong and efficient network structure that facilitate efficient product recommendation and are robust against random failures however, weak against targeted attacks on said hubs.

The roles of the products in the network were also determined by analysing the in-degree and out-degree distributions. Products with a high in-degree are typically bought alongside others and can used to evaluate the products popularity or demand within the network. High in-degree products are prioritised in the recommender system as they are suggested to have a broader appeal or compatibility with different products as they often align with customers interest which results in them being frequently recommended and bought with other products. Opposingly, products with a high out-degree are co-purchased with other products and play an important role in influencing purchasing patterns by introducing customers to different product categories.

The average shortest path length metric was calculated using NetworkX. The result shows a reading of an overall average path length of 4.226 rounded to 3 decimal points, which in comparison to the size of the network is a low value. This network property quantifies how connected the network is in terms of the average number of steps it takes to get from a random product to another through co-purchasing connections. A low average shortest path length would indicate high navigability with intricately connected products in a network. In relation to the Amazon co-purchasing network, this would suggest that an efficient product network where customers are able to find a diverse range of new products with few co-purchases. This property is key in improving the recommendation system, prioritising the recommendations that require the fewest steps possible.

The clustering coefficient evaluates the degree in which the nodes are clustered together in the network. With Gephi, we were able to calculate this value to be 0.360 which is relatively high compared to the density of the network. This metric informs us regarding the frequency of which co-purchased products of any given product are also bought together, indicating the closeness of their relationship in a network. Understanding this measure allows us to identify product categories and improve product recommendations through the analysis of relationships. Product in an area of high clustering might be recommended with each other more frequently while product with product linking clusters could help introduce customers to new product categories.



Determining the type of network, specifically whether the Amazon co-purchasing network is a random or a small-world network, is crucial in understanding how each product is connected and the efficiency of the network's topology. Before exploring the specific properties that define the type of the network, it is important to recognize that while we have previously highlighted these defining key measures, we lack a basis of comparison to definitively identify the type of network. Identifying this will aid in contextualising these metrics into a foundation that can explain the behaviour of the recommendation system and its strengths and weaknesses in terms of information flow within the network. This plays a crucial role in user experience on the website, addressing how customers can navigate Amazon's product network, and influencing the effectiveness of the product recommendation system and the discovery of new products. Following this, by using the Erdős-Rényi model, we can compare the average shortest path length and clustering coefficient that we have calculated for our network against the expected values for a random network of the same size to identify the network type and its impact on consumer behaviour. Small-world networks, a concept first explored by Stanley Milgram and later fully developed by Duncan Watts and Steven Storgatz, are a class of networks that are characterized by their uniquely high clustering yet short average path length [9]. These two properties result in a network with efficient diffusion of information despite having a high degree of interconnectedness community structure within the network. Small-world networks are often used to effectively model real-world networks to facilitate robust and efficient systems.

Opposingly, Random networks are networks that are constructed by randomly connecting nodes without thought for distance and preferences between each other. This erratic creation pattern leads to properties such as low clustering and variable path lengths, thus these networks have a lack of cohesive groups and community structure, inefficient for process that rely on closely related groups of nodes. Overall, this means that random networks have limited real world representations, therefore are used as a theoretical benchmark to analyse network structure and properties through contrast of comparison.

To determine the classification of the Amazon's Product Co-purchasing network, we will have to compare this network to a random network with the same size and properties. This can be done by employing the Erdős-Rényi Model to generate expected the expected clustering coefficient and average shortest path length metrics for a random undirected network graph to serve as our basis for comparison.

Before we can calculate the expected shortest path length in a random network using the Erdős–Rényi (E-R) graph, we first must calculate —the average clustering coefficient ***p***. This can be done by using the equation seen below where ***m*** is the total number of number of edges and ***n*** is the total number of nodes in the network. With the Amazon co-purchasing network having **10,001 nodes** and **41,763 edges**, the probability of any two nodes being connected to an edge equates to **4.176×10^{-4}**.

$$p = \frac{m}{n(n-1)}$$

With calculated, the expected average shortest path length can be calculated using the equation seen in the formula below with the same notations used in the formula for calculating ***p***. We found that the expected shortest path length of this random network is **6.443**.

$$\mathbb{E}[L] \approx \frac{\log(n)}{\log(np)}$$

Comparing the metrics between the expected random network and the Amazon Product Co-Purchasing Network, we found that Amazon's network has a shorter average shortest path length with a length of **4.226** compared to the random network's **6.443**. As for clustering coefficients, Amazon's network has a higher clustering coefficient of **0.360** compared to the random network's **4.176×10^{-4}**. In context of small world networks, the Amazon Product Co-Purchasing Network exhibits properties typical to small world networks, with short average path lengths and high clustering coefficients.

Establishing that the Amazon network demonstrates attributes of a small world network, we identify several implications to the network's functionality, efficiency, and vulnerabilities. The network having a short path length implies quick distribution of information, crucial for spreading product recommendations and reviews to users. High clustering coefficients also inform us that the network can facilitate the development of reliable recommendation systems.

Establishing that the Amazon network demonstrates attributes of a small world network, we identify several implications to the network's functionality, efficiency, and vulnerabilities. The network having a short path length implies quick distribution of information, crucial for spreading product recommendations and reviews to users. High clustering coefficients also inform us that the network can facilitate the development of reliable recommendation systems that can suggest relevant products to potential customers, driving sales. These networks also tend to form strong community structure that can be used to define product categories for targeted marketing.

While the Amazon network does gain many advantages from its small world structure, it also faces vulnerabilities including potential attacks to the main hubs and communities, rapid spread of harmful content and reduced exposures to niche product groups. With highly connected products and product categories playing a critical role of maintaining the whole networks connectivity, these hubs are exposed to malicious attacks, system failures, or inventory issues that could disrupt the entirety of the network. Should this error inflict the network, it could cause these hubs to disappear, leading to the network fragmenting and a significant reduction the flow of information, thereby hindering the effectiveness of the system.

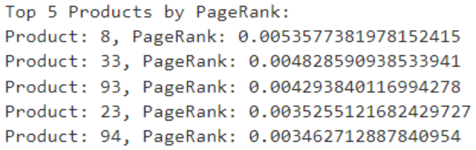
The speed and efficiency that allows for fast distribution of information seen in a small world network could also be detrimental to the safety of the users and the network as it could allow for the rapid spreading of harmful content. Such content could include misinformation, negative reviews or sales of potentially harmful products. The high degree of connectivity and low average path length means that such information could quickly spread to reach a large population, leading to stockouts, brand damage or advertising of a bad product.

Tight knit communities, a prominent characteristic of small world networks, contribute to a large part of creating specialised categories and improving customer experience. While this may be beneficial for target marketing, it also runs the risk of isolating users to products that they are already exposed to, making the discovery of new products difficult. This reduces the effectiveness of Amazons product marketing by reinforcing existing user preferences rather than allowing users to explore new products.

These issues inherent to small-world networks can be addressed through an approach that promotes diversity and improves strength between connections in the network. To fix the issue, Amazon should have real time monitoring and detection systems to prevent the fast spread of harmful content and decentralise the network that that depends less on hubs and more on smaller clusters, distributing the network's critical point evenly throughout the network. This can be achieved by slightly decreasing the overall clustering coefficient through diversifying the product recommendation, adjusting network algorithms and allow for customisable user recommendations. The addition of features like customisable user recommendations informs users about what they are recommended and allows users to adjust their preferences to explore new product categories they are interested in rather than being force fed the same recommendations repeatedly. This feature would introduce new paths between existing products and product categories, thereby slowly decentralising the network over time through diversifying user interaction.

Our study also utilises the PageRank algorithm, which was initially created for ranking websites in search engine queries but is also effective at revealing the significance of nodes inside a network, to explore the complex web of Amazon's product co-purchasing network. The PageRank algorithm, developed by Brin and Page, is based on the concept that the importance of a web page can be determined by the number of hyperlinks pointing to it [1]. Each node in this network analysis represents a product, and co-purchasing relationships are shown by directed edges. The PageRank algorithm sets itself in a unique position to find important products that have a significant impact on the dynamics of the network because of their extensive co-purchase relationships.

Using the NetworkX python library, we implemented the PageRank algorithm, utilising its effective calculations to provide importance scores to products according to their connectivity and the network's overall structure. We followed the default configuration, which included a damping factor of 0.85, which represents the likelihood that a client will keep scrolling through similar product recommendations.



```
Top 5 Products by PageRank:
Product: 8, PageRank: 0.0053577381978152415
Product: 33, PageRank: 0.004828590938533941
Product: 93, PageRank: 0.004293840116994278
Product: 23, PageRank: 0.0035255121682429727
Product: 94, PageRank: 0.003462712887840954
```

The generated PageRank scores provided an ordered list of products inside the network from most to least influential, allowing for a focused analysis of the products that are most significant regarding customers' purchase patterns.

VI. Conclusion

This study revealed that Amazon's co-purchasing network has small-world traits that improve suggestion efficiency but introduce vulnerabilities because of reliance on a central hub. Our use of Gephi and NetworkX, allowed us to identify important items and community structures, which helped us recommend modifications to our marketing plans. Future research should aim to reduce network dependencies and introduce more diverse recommendations to minimise bias and enhance resilience. Amazon's customer interactions and market performance can be further optimised by modifying methods in response to network dynamics.

References

[1]
M. Henzinger, “PageRank Algorithm,” *Springer eBooks*, pp. 1509–1511, Jan. 2016, doi: https://doi.org/10.1007/978-1-4939-2864-4_277.

[2]
M. S. Ahuja, “USING COMMUNITY DETECTION TECHNIQUE IN RECOMMENDER SYSTEM,” *Advances in Mathematics: Scientific Journal*, vol. 9, no. 6, pp. 3741–3750, Jul. 2020, doi: https://doi.org/10.37418/amsj.9.6.52.

[3]
M. Kumbhar, J. Kolhe, D. Kumawat, and S. P. Bansu, "Employing Community Detection into Recommender System: A Review," 2020 Available: https://api.semanticscholar.org/CorpusID:235816047

[4]
G. L. Li, H. Li, Y. R. Wang, and T. B. Zhang, “The Solution to Node Importance in Complex Networks Based on PageRank Algorithm,” *Applied Mechanics and Materials*, vol. 599–601, pp. 1777–1780, Aug. 2014, doi: https://doi.org/10.4028/www.scientific.net/amm.599-601.1777.

[5]
X. F. Meng, "BRING PageRank TO THE INFRASTRUCTURE NETWORK," 2018. . Available: https://api.semanticscholar.org/CorpusID:199420607

[6]
Liu, Y., Wu, C. and Tong, X. *Prediction of Co-purchasing Product*, *cseweb*. Available at: https://cseweb.ucsd.edu/classes/sp15/cse190-c/reports/sp15/039.pdf.

[7]
Prasad, U. *et al.* (2017) *Analysis of the co-purchase network of products to predict Amazon sales-rank*, *SpringerLink*. Available at: https://link.springer.com/chapter/10.1007/978-3-319-72413-3_13 (Accessed: 10 April 2024).

[8]
Basuchowdhuri, P., Shekhawat, M.K. and Saha, S.K. (2015) *Analysis of Product Purchase Patterns in a Co-Purchase Network*, *ieeexplore*. Available at: https://ieeexplore.ieee.org/abstract/document/7052071/authors#authors (Accessed: 10 April 2024).

[9]

Q. K. Telesford, K. E. Joyce, S. Hayasaka, J. H. Burdette, and P. J. Laurienti, (2011) “The ubiquity of small-world networks," Brain Connect., vol. 1, no. 5, pp. 367-375. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3604768/ (Accessed: 7 April 2024).