Social Network Analysis of Fine Food Reviewers

Dataset Introduction

The Amazon Fine Foods Reviews data set is a text file that contains 568,454 records of user reviews of fine food products from Amazon. Each record contains information on the user ID, product ID, profile name, helpfulness, score, time, summary and text, which can be seen in Figure 1.

Network Structure

This project aims to construct and analyze a network of Amazon fine food reviewers based on the similarity of their product reviews, in other words, the goal is to detect communities of users who have similar tastes or opinions for fine food products on Amazon. Each node will represent a user and edges between the nodes will exist if two users gave the same product the same score in their review. The weights for each edge represent the number of similar reviews each pair of users has. Due to the large size of the dataset, the project will focus on the relationship between the top 50 reviewers, which is based on the number of reviews recorded for each user in the dataset. Figure 2 shows a dataframe of the graph data.

Basic Network Statistics

After constructing the network, there are a total of 48 nodes in the network with 461 edges between the nodes. Given that the original dataset had records for 50 users, this indicates that two of the top 50 reviewers had no similar reviews to any of the other top 50 reviewers. On average for each reviewer, about 9 other reviewers in the top 50 have also given a product a review with the same score.

The degree of a node measures the number of neighbours for that node, where neighbours are the other nodes that are connected to the node by an edge. Therefore, in this case, the degree of a node will return the number of reviewers who have given the same product the same score. Figure 3 in the Appendix shows the degree for each node in the network. We can see that user ‘AKMEY1BSHSDG7’ has the highest degree, with connections to 36 other users who have written at least one similar review to this user. At the bottom of the figure, we can see that user ‘AKZKG2Z7CNV27’ only has one review that is similar to another user in the top 50 reviewers.

Since the network is a weighted graph, we can also compute the weighted degree of each node, where the weighted degree is calculated as the sum of the weighted edges connected to the node (NetworkX 3.2.1 documentation, n.d.). Compared to the unweighted degree, the weighted degree not only factors in the number of connections to the node but also the strength of the connection

measured by the weight of the edge connecting the two nodes. In Figure 4, we can see that the node with the highest weighted degree is now user ‘A1TMAVN4CEM8U8’ with a weighted degree of 461. Given that the weight of a node is the number of similar reviews a user has with other users, we can say that user ‘A1TMAVN4CEM8U8’ has scored a product the same as another user 461 times.

By comparing the two measures, we can say that user ‘AKMEY1BSHSDG7’, who had the highest degree, has the highest number of connections, whereas user ‘A1TMAVN4CEM8U8’, with the highest weighted degree, had a stronger weighted connection to its neighbours. In other words, the second user had a higher number of similar opinions to each of its neighbouring reviewers whereas the first user simply had more similar reviews to a higher number of distinct reviewers.

Shortest Paths

Given this network, we can also find the shortest path between two reviewers. For example, we can first look at the shortest path between the reviewer with the highest simple degree, ‘AKMEY1BSHSDG7’, and the reviewer with the second highest simple degree, 'AQLL2R1PPR46X'. Using NetworkX’s functions, shortest_path() and shortest_path_length(), we found that these two users are directly connected by one edge, resulting in a shortest path with a length of 1. On the other hand, the shortest path between the reviewer with the highest simple degree and the reviewer with the lowest simple degree has a length of 2, with user ‘A36MP37DITBU6F’ connecting the two other reviewers.

Node Importance

To better understand the influence of each node in the network, we can look at various measures of node importance such as Degree Centrality, Betweenness Centrality, and Closeness Centrality.

Degree Centrality measures the fraction of total nodes a node is connected to, or how connected the node is to the rest of the network (NetworkX 3.2.1 documentation, n.d.). In Figure 5, we can see that user ‘AKMEY1BSHSDG7’ had the highest degree centrality at about 0.77, which indicates that this user had written at least one similar review to 77% of the top 50 Amazon Users. This is also the same user with the highest degree in Figure II.

The Betweenness Centrality measures how important a node is in terms of connecting other nodes to each other (Neo4j Graph Data Platform, (n.d.-a). Specifically, the Betweenness Centrality of node V is the fraction of the shortest paths from node A to node B that contain node V over the total number of shortest paths connecting node A to node B (Padmanabhan, 2023). In Figure 6, we can see that all nodes in the network have a very low betweenness centrality score, with user ‘A36MP37DITBU6F’ having the highest betweenness centrality at only 0.059. This is expected as edges are created for each pair of nodes if the two reviewers share a similar review

for a product. Therefore in this network, most pairs of nodes will be directly connected with a shortest path containing only 1 edge, resulting in a very low betweenness centrality for every node.

Finally, the Closeness Centrality measures how far a node is from all the other nodes, with a higher closeness centrality score corresponding to nodes that are closer to all other nodes (Neo4j Graph Data Platform, (n.d.-b). In Figure 7, we can see that user 'AKMEY1BSHSDG7' has the highest closeness centrality score at 0.81; this is the same user with the highest simple degree and degree centrality score. Since a lot of the shortest paths in the network contain only one edge, this user scored the highest in closeness centrality as it had the most neighbours in this network, consequently resulting in shorter distances to all other nodes in the network.

Community Detection

After constructing the network we can try to identify communities within the network using Louvain Modularity. Louvain Modularity identifies communities through optimization based on modularity, where modularity compares the density of connections within clusters and the density of connections between clusters (Padmanabhan, 2023). After running the Louvain Modularity algorithm, the output showed that the algorithm was able to identify 6 communities in this network. Figure 8 shows a plot of the network, with the different colours identifying the different communities.

Community Attributes

To better understand the characteristics of each community, we can analyze the similar reviews between each pair of reviewers in each community. After gathering similar reviews between each pair of reviewers in each community, we can plot the distribution of scores given by the reviewers in each community. Figure 9 shows the average score for each community and Figure 9 shows the bar plots of the number of reviews with each score for each community. From these figures, we can see that the majority of communities have very high scores for their reviews and that Communities 0 and 4 only have reviews with a score of 5. Only Community 2 has reviews with scores ranging from 1-5. This implies that many of the reviewers in each community share positive opinions on products if a high score corresponds to a positive opinion.

Given that the dataset did not include the product name for each review, we can analyze the summary text and review text to better understand what kind of product the user reviewed. To do so, we can create a word cloud of the top 100 words found for each community's review summaries and review texts. Figures 11-16 show the word clouds generated for each community.

Out of the top words generated for community 0, we can see that this community contains users who shared positive reviews on various snacks such as chips, chocolate, and cereal, describing them with words such as 'crunch' or 'hot'. Community 1 also seems to have reviewers who

mainly reviewed products such as chips, but also other products such as tea, beef, fruit, and juice. Communities 2 seems to focus more on tea reviews, commonly using words such as ‘sweet’ and ‘organic’. Community 3 also focuses on tea reviews, but particularly types like Chamomile and hibiscus. Community 4 has a variety of food reviews including almonds and pasta, with common descriptive words such as ‘bold’ and ‘spicy’. Finally, Community 5 has a strong emphasis on coffee reviews, with descriptive words such as ‘bold’, ‘roast’ and ‘acidic’.

Conclusion

From this project, we were able to construct and analyze reviewers in the Amazon fine foods dataset. To improve on this network in the future, we can try to re-define the edges of the network to better measure the similarity between reviewers. For example, we can take the review text into consideration to measure the similarity of reviews between two users.

References

Betweenness centrality - neo4j graph data science. Neo4j Graph Data Platform. (n.d.-a). https://neo4j.com/docs/graph-data-science/current/algorithms/betweenness-centrality/

Closeness centrality - neo4j graph data science. Neo4j Graph Data Platform. (n.d.-b). https://neo4j.com/docs/graph-data-science/current/algorithms/closeness-centrality/#algorithms-closeness-centrality-limitations

Degree_centrality. degree_centrality - NetworkX 3.2.1 documentation. (n.d.). https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.degree_centrality.html

Graph.degree. Graph.degree - NetworkX 3.2.1 documentation. (n.d.). https://networkx.org/documentation/stable/reference/classes/generated/networkx.Graph.degree.html

Padmanabhan, K. (2023a, December). *Social Network Analysis - Week 1*. Lecture.

Padmanabhan, K. (2023b, December). *Social Network Analysis - Week 2*. Lecture.

Appendix

Figure 1: Amazon Fine Foods Dataset Sample

	ProductId	UserId	ProfileName	Helpfulness	Score	Time	Summary	Text
0	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1/1	5.0	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0/0	1.0	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1/1	4.0	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	B000UA0QIQ	A395BORC6FGVXV	Karl	3/3	2.0	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0/0	5.0	1350777600	Great taffy	Great taffy at a great price. There was a wid...

Figure 2: Weighted Network Graph Data

	User A	User B	Weight
1	A3FKGKUCI3DG9U	AEC90GPFKLAAW	2
6	A2FRFAQCWZJT3Q	AY1EF0GOH80EK	1
9	A25C2M3QF9G7OQ	A2SZLNSI5KOQJT	5
11	A1IU7S4HCK1XK0	A2FRFAQCWZJT3Q	1
12	A1WX42M589VAMQ	A2DPYMNI2HCIOI	1
...	...	...	...
1214	AQLL2R1PPR46X	AY12DBB0U420B	1
1215	A250AXLRBVYKB4	A2FRFAQCWZJT3Q	11
1216	A1X1CEGHTHMBL1	A2R6RA8FRBS608	1
1219	A1P2XYD265YE21	A36MP37DITBU6F	3
1224	A2SZLNSI5KOQJT	A3FKGKUCI3DG9U	2

461 rows × 3 columns

Figure 3: Simple Degree

```
 [('AKMEY1BSHSDG7', 36),
  ('A3OXHLG6DIBRW8', 32),
  ('AQLL2R1PPR46X', 32),
  ('A3PJZ8TU8FDQ1K', 30),
  ('A2FRFAQCWZJT3Q', 28),
  ('A25C2M3QF9G7OQ', 28),
```

Figure 4: Weighted Degree

```
[('A1TMAVN4CEM8U8', 461),
 ('ADS5APY1NKTL4', 421),
 ('A1X1CEGHTHMBL1', 396),
 ('A3PJZ8TU8FDQ1K', 375),
 ('A281NPSIMI1C2R', 256),
 ('A3OXHLG6DIBRW8', 249),
 ('A3FY3H6F4249E0', 242),
 ('A2FRFAQCWZJT3Q', 226),
 ('AKMEY1BSHSDG7', 219),
 ('A1YUL9PCJR3JTY', 208),
 ('A1P2XYD265YE21', 197),
 ('A1LZJZIHUPLDV4', 188),
 ('AY1EF0GOH80EK', 167),
 ('A2Y8IDC1FKGNJC', 167),
 ('AQLL2R1PPR46X', 146),
 ('A1WX42M589VAMQ', 143),
 ('A3FKGKUCI3DG9U', 141),
 ('A25C2M3QF9G7OQ', 129),
 ('A2SZLNSI5KOQJT', 122),
 ('A1Z54EM24Y40LL', 115),
 ('A1HRYC60VTMYC0', 114),
 ('AQQLWCMRNDFGI', 112),
 ('A1UQBFCERIP7VJ', 103),
 ('A2PNOU7NXB1JE4', 96),
 ('A3QNQQKJTL76H0', 95),
 ('A17HMM1M7T9PJ1', 90),
 ('A250AXLRBVYKB4', 71),
 ('A3HPCRD9RX351S', 61),
 ('A31N6KB160O508', 60),
 ('A1IU7S4HCK1XK0', 54),
 ('A2MUGFV2TDQ47K', 52),
 ('A1ZH9LWMX5UCFJ', 51),
 ('AEC90GPFKLAAW', 49),
 ('A36MP37DITBU6F', 48),
 ('A2XNOB1T796Y6B', 44),
 ('ALSAOZ1V546VT', 40),
 ('A2DPYMNI2HCIOI', 35),
 ('AY12DBB0U420B', 35),
 ('A2R6RA8FRBS608', 31),
 ('AZV26LP92E6WU', 30),
 ('A3F3B1JPACN215', 28),
 ('A33AQPJYH7UUXR', 26),
 ('A35R32TA60XD57', 21),
 ('A2M9D9BDHONV3Y', 21),
 ('A36WGHR8TO5DKT', 19),
 ('A2GEZJHBV92EVR', 14),
 ('A3D6OI36USYOU1', 5),
 ('AKZKG2Z7CNV27', 1)]
```

Figure 5: Degree Centrality

```
[('AKMEY1BSHSDG7', 0.7659574468085106),
 ('A3OXHLG6DIBRW8', 0.6808510638297872),
 ('AQLL2R1PPR46X', 0.6808510638297872),
 ('A3PJZ8TU8FDQ1K', 0.6382978723404256),
 ('A2FRFAQCWZJT3Q', 0.5957446808510638),
 ('A25C2M3QF9G7OQ', 0.5957446808510638),
 ('A2MUGFV2TDQ47K', 0.5957446808510638),
 ('A2Y8IDC1FKGNJC', 0.5957446808510638),
 ('A281NPSIMI1C2R', 0.5957446808510638),
 ('A1P2XYD265YE21', 0.5744680851063829),
 ('A3FKGKUCI3DG9U', 0.5531914893617021),
 ('A1HRYC60VTMYC0', 0.5531914893617021),
 ('AY1EF0GOH80EK', 0.5106382978723404),
 ('A1IU7S4HCK1XK0', 0.5106382978723404),
 ('A1UQBFCERIP7VJ', 0.5106382978723404),
 ('ALSAOZ1V546VT', 0.48936170212765956),
 ('A3HPCRD9RX351S', 0.48936170212765956),
 ('AEC90GPFKLAAW', 0.46808510638297873),
 ('A36MP37DITBU6F', 0.46808510638297873),
 ('A3QNQQKJTL76H0', 0.46808510638297873),
 ('A17HMM1M7T9PJ1', 0.44680851063829785),
 ('A1Z54EM24Y40LL', 0.425531914893617),
 ('A2SZLNSI5KOQJT', 0.40425531914893614),
 ('A2PNOU7NXB1JE4', 0.40425531914893614),
 ('A2DPYMNI2HCIOI', 0.3829787234042553),
 ('A3F3B1JPACN215', 0.3829787234042553),
 ('ADS5APY1NKTL4', 0.3829787234042553),
 ('A2XNOB1T796Y6B', 0.3829787234042553),
 ('A250AXLRBVYKB4', 0.3829787234042553),
 ('A1ZH9LWMX5UCFJ', 0.36170212765957444),
 ('A1LZJZIHUPLDV4', 0.36170212765957444),
 ('A1YUL9PCJR3JTY', 0.36170212765957444),
 ('AY12DBB0U420B', 0.36170212765957444),
 ('A2R6RA8FRBS608', 0.3404255319148936),
 ('A1TMAVN4CEM8U8', 0.3191489361702128),
 ('A3FY3H6F4249E0', 0.3191489361702128),
 ('AZV26LP92E6WU', 0.3191489361702128),
 ('A2M9D9BDHONV3Y', 0.3191489361702128),
 ('A33AQPJYH7UUXR', 0.2765957446808511),
 ('A1WX42M589VAMQ', 0.23404255319148937),
 ('A1X1CEGHTHMBL1', 0.23404255319148937),
 ('A2GEZJHBV92EVR', 0.2127659574468085),
 ('AQQLWCMRNDFGI', 0.19148936170212766),
 ('A31N6KB160O508', 0.1702127659574468),
 ('A36WGHR8TO5DKT', 0.1276595744680851),
 ('A35R32TA60XD57', 0.0851063829787234),
 ('A3D6OI36USYOU1', 0.06382978723404255),
 ('AKZKG2Z7CNV27', 0.02127659574468085)]
```

Figure 6: Degree Betweenness

```
[('A36MP37DITBU6F', 0.05861583093828492),
 ('AKMEY1BSHSDG7', 0.05279686190126074),
 ('A3OXHLG6DIBRW8', 0.04174389176771267),
 ('A2MUGFV2TDQ47K', 0.035192673374750946),
 ('A3PJZ8TU8FDQ1K', 0.029574230918795654),
 ('A2FRFAQCWZJT3Q', 0.02721618556339733),
 ('A281NPSIMI1C2R', 0.02442729753888083),
 ('A1HRYC60VTMYC0', 0.022889721688026488),
 ('A1UQBFCERIP7VJ', 0.02285299122558476 2),
 ('AQLL2R1PPR46X', 0.02260657615350786),
 ('A1P2XYD265YE21', 0.01899911986118142),
 ('AY1EF0GOH80EK', 0.0172374410025772),
 ('A2Y8IDC1FKGNJC', 0.016922068736142156),
 ('A3HPCRD9RX351S', 0.016243134298401463),
 ('A25C2M3QF9G7OQ', 0.015350665566395754),
 ('ALSAOZ1V546VT', 0.014669497567233743),
 ('A1IU7S4HCK1XK0', 0.013553717760456444),
 ('A3FKGKUCI3DG9U', 0.013230970657261896),
 ('A3QNQQKJTL76H0', 0.012890820705938576),
 ('A2SZLNSI5KOQJT', 0.012611858872 36256),
 ('A1Z54EM24Y40LL', 0.012529069828911898),
 ('A250AXLRBVYKB4', 0.0108380641412553),
 ('A2R6RA8FRBS608', 0.009677491589670528),
 ('A2DPYMNI2HCIOI', 0.00924588505585776),
 ('A2M9D9BDHONV3Y', 0.008654213527009328),
 ('AY12DBB0U420B', 0.008547595492510206),
 ('A3F3B1JPACN215', 0.008336520796320565),
 ('A2PNOU7NXB1JE4', 0.008069398169992305),
 ('AEC90GPFKLAAW', 0.007932367608579008),
 ('A2XNOB1T796Y6B', 0.007269410433770456),
 ('A17HMM1M7T9PJ1', 0.006911872172509754),
 ('AZV26LP92E6WU', 0.006164479698515761),
 ('A1ZH9LWMX5UCFJ', 0.006087788237350189),
 ('A31N6KB160O508', 0.005641107606880039 5),
 ('A1LZJZIHUPLDV4', 0.005613325386683 3864),
 ('A2GEZJHBV92EVR', 0.005529003206881639),
 ('A1YUL9PCJR3JTY', 0.005176828925343 0536),
 ('AQQLWCMRNDFGI', 0.00498209892671023 1),
 ('A1X1CEGHTHMBL1', 0.004867720173951983),
 ('A1TMAVN4CEM8U8', 0.004658491625122907),
 ('A33AQPJYH7UUXR', 0.003916542708775849),
 ('ADS5APY1NKTL4', 0.003896686629228 1014),
 ('A3FY3H6F4249E0', 0.003865625936012833),
 ('A1WX42M589VAMQ', 0.003699363643859481),
 ('A36WGHR8TO5DKT', 0.002167051953834743),
 ('A35R32TA60XD57', 0.000120492230 1945755),
 ('A3D6OI36USYOU1', 0.0),
 ('AKZKG2Z7CNV27', 0.0)]
```

Figure 7: Closeness Centrality

```
[('AKMEY1BSHSDG7', 0.8103448275862069),
 ('A3OXHLG6DIBRW8', 0.7580645161290323),
 ('AQLL2R1PPR46X', 0.7580645161290323),
 ('A3PJZ8TU8FDQ1K', 0.734375),
 ('A2FRFAQCWZJT3Q', 0.7121212121212122),
 ('A2MUGFV2TDQ47K', 0.7121212121212122),
 ('A2Y8IDC1FKGNJC', 0.7121212121212122),
 ('A25C2M3QF9G7OQ', 0.7014925373134329),
 ('A1P2XYD265YE21', 0.7014925373134329),
 ('A281NPSIMI1C2R', 0.7014925373134329),
 ('A1HRYC60VTMYC0', 0.6911764705882353),
 ('A3FKGKUCI3DG9U', 0.6811594202898551),
 ('AY1EF0GOH80EK', 0.6714285714285714),
 ('A1IU7S4HCK1XK0', 0.6619718309859155),
 ('ALSAOZ1V546VT', 0.6619718309859155),
 ('A3HPCRD9RX351S', 0.6619718309859155),
 ('A1UQBFCERIP7VJ', 0.6619718309859155),
 ('AEC90GPFKLAAW', 0.6527777777777778),
 ('A36MP37DITBU6F', 0.6527777777777778),
 ('A3QNQQKJTL76H0', 0.6438356164383562),
 ('A17HMM1M7T9PJ1', 0.6438356164383562),
 ('A1Z54EM24Y40LL', 0.6351351351351351),
 ('A2SZLNSI5KOQJT', 0.6266666666666667),
 ('A2DPYMNI2HCIOI', 0.618421052631579),
 ('A2PNOU7NXB1JE4', 0.618421052631579),
 ('A1LZJZIHUPLDV4', 0.6103896103896104),
 ('A3F3B1JPACN215', 0.6103896103896104),
 ('ADS5APY1NKTL4', 0.6103896103896104),
 ('A2XNOB1T796Y6B', 0.6103896103896104),
 ('A1ZH9LWMX5UCFJ', 0.6025641025641025),
 ('AY12DBB0U420B', 0.6025641025641025),
 ('A250AXLRBVYKB4', 0.6025641025641025),
 ('A1YUL9PCJR3JTY', 0.5949367088607594),
 ('A2R6RA8FRBS608', 0.5875),
 ('AZV26LP92E6WU', 0.5875),
 ('A1TMAVN4CEM8U8', 0.5802469135802469),
 ('A2M9D9BDHONV3Y', 0.5802469135802469),
 ('A33AQPJYH7UUXR', 0.573170731707317),
 ('A3FY3H6F4249E0', 0.573170731707317),
 ('A1WX42M589VAMQ', 0.5595238095238095),
 ('A1X1CEGHTHMBL1', 0.5529411764705883),
 ('A2GEZJHBV92EVR', 0.5529411764705883),
 ('AQQLWCMRNDFGI', 0.5402298850574713),
 ('A36WGHR8TO5DKT', 0.5222222222222223),
 ('A31N6KB160O508', 0.5222222222222223),
 ('A3D6OI36USYOU1', 0.47959183673469385),
 ('A35R32TA60XD57', 0.4747474747474475),
 ('AKZKG2Z7CNV27', 0.3983050847457627)]
```

Figure 8: Community Identification

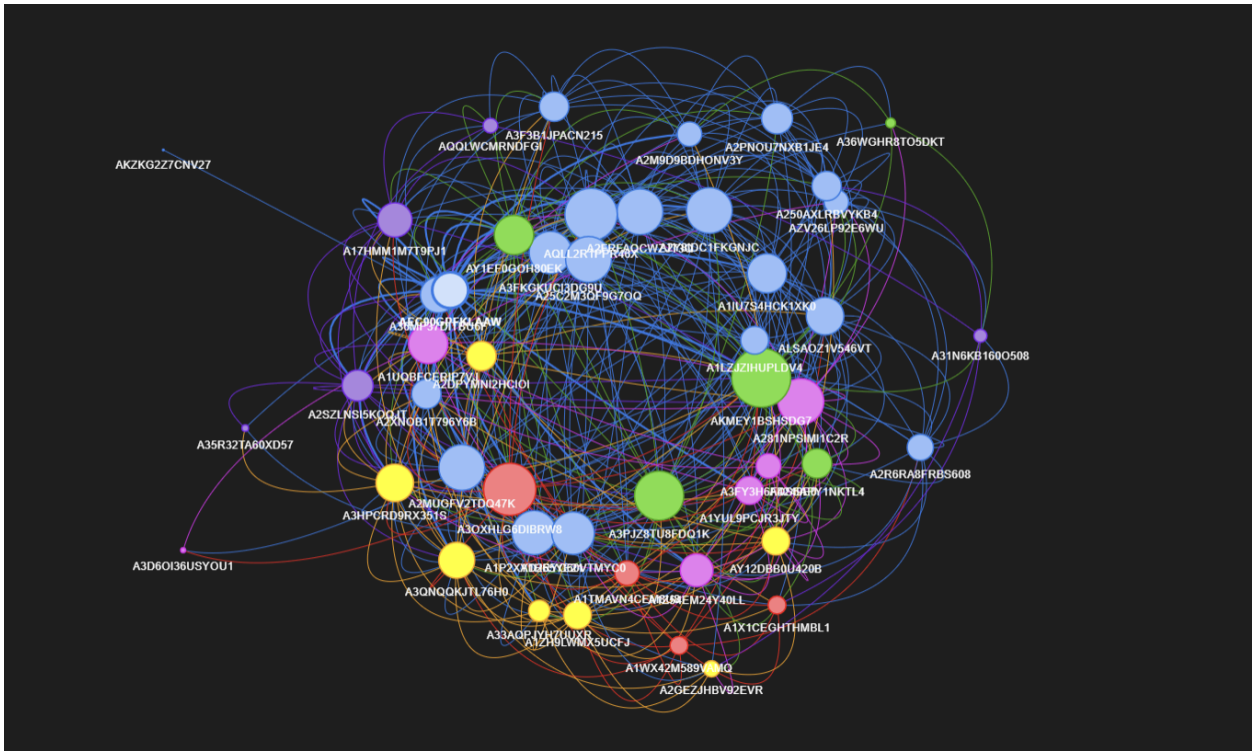

Figure 9: Average Review Score by Community

```
Community
0    5.000000
1    4.614035
2    4.554723
3    4.829352
4    5.000000
5    4.148148
```

Figure 10: Review Scores for Each Community

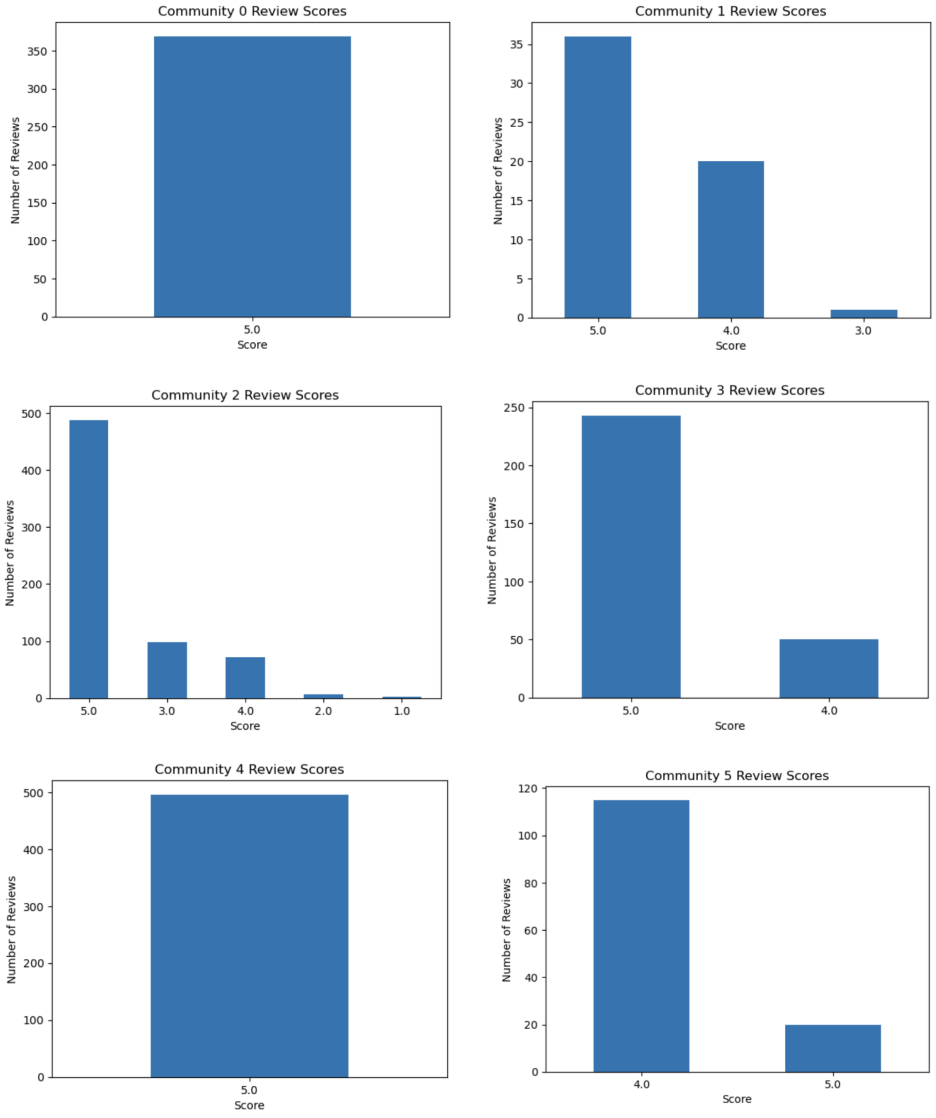

Figure 11: Word Clouds for Cluster 0 (Left: Summary text, Right: Review Text)

Figure 12: Word Clouds for Cluster 1 (Left: Summary text, Right: Review Text)

Figure 13: Word Clouds for Cluster 2 (Left: Summary text, Right: Review Text)

Figure 14: Word Clouds for Cluster 3 (Left: Summary text, Right: Review Text)

Figure 15: Word Clouds for Cluster 4 (Left: Summary text, Right: Review Text)

Figure 16: Word Clouds for Cluster 5 (Left: Summary text, Right: Review Text)