Nikita Salkar DS210 Final Project Report

Collaborators: Om Italiya, Sia Sharma and Ananya Agarwal. We helped each other debug and worked on collating a holistic approach towards tackling the project.

**Introduction**

My project looks to analyze a dataset from the “Arxiv GR-QC collaborations network,” which covers papers submitted to the General Relativity and Quantum Cosmology category on Arxiv. The dataset maps scientific collaborations between authors, representing them as an undirected graph. In this graph, nodes represent individual authors, and edges indicate co-authorships between authors. An edge between nodes ‘I’ and ‘j’ have co-authored at least one paper together. Papers with multiple authors create fully connected subgraphs, capturing all possible co-authorship relationships.

**Problem Context**

My analysis aims to explore the collaboration patterns and connectivity within the GR-QC research network, addressing key questions such as: Are friends of friends truly connected? This examines the similarity between co-author networks of pairs and nodes with the shortest path of 2, indicating they share a mutual co-author. How strong are the collaborations in the GR-QC network? This explores the similarity between the networks of co-authors of pairs of nodes with the shortest path of 2, highlighting the potential for deeper collaboration. Lastly, what is the overall structure of the GR-QC network? This analysis provides statistical insights into the mean, maximum, and distribution of Jaccard similarity scores across the network, uncovering patterns and areas with potential for stronger connections.

**Solution Approach**

To address these important questions, the solution employs several algorithms and metrics. The dataset is read into a graph stored as an adjacency list in a ‘HashMap’. This structure facilitates efficient computations of various metrics and allows for the exploring of the network’s connectivity. I used Breadth-First Search (BFS) to compute the shortest path between nodes, identifying pairs with the shortest path of 2 for further analysis. Jaccard Similarity compares the sets of neighbors between pairs of nodes with the shortest path of 2, yielding a score ranging from 0 to 1. The intersection of these sets represents the shared neighbors between the two nodes, while the union represents all distinct neighbors across both nodes. The Jaccard similarity score is the ratio of the intersections size to the union size, providing a measure of similarity between authors’ co-authorship networks.

The mean and maximum Jaccard similarity scores are computed, along with pairs achieving the highest scores, offering insights into the network’s overall collaboration patterns, and highlighting strong connections. Additionally, the percentage of pairs exceeding thresholds ranging from 0.1 to 1.0 in 0.1 increments is calculated, illustrating the distribution of collaboration strengths across the network.

**Jaccard Similarity**

Jaccard Similarity is a metric that measures the similarity between two sets, defined as the ratio between the size of their intersection and the size of their union. In the context of my project, it compares the co-author networks of pairs of authors in the GR-QC field. This provides insight into how similar or distinct their collaboration patterns are.

A score of 1.00 indicates that the two authors have identical co-author networks, while a score closer to 0.00 suggests little or no overlap between their networks. This metric offers a straightforward way to quantify the degree of connection between authors’ collaboration networks and can reveal patterns in the GR-QC network, such as tightly knit research groups or more isolated authors.

**Implementation**

My project code comprises 3 main modules. ‘graph.rs’ constructs the graph by reading the dataset into an adjacency list stored in a ‘HashMap’. It also provides functions to compute the BFS shortest path between nodes, returning a ‘HashMap’ of distances, and retrieve the neighbors of a given node. ‘Jaccard\_similarity.rs’ calculates the Jaccard similarity scores for pairs of nodes with a shortest path of 2, storing them in a ‘HashMap’, and compares the neighbors of pairs of nodes, computing the intersection and union to derive the similarity score. Within this module I also have a function that computes the mean, maximum, and associated pairs of Jaccard similarity scores, and calculates the percentage of pairs above specified thresholds, offering insights into the distribution of collaboration strengths. The ‘main.rs’ integrates all the modules, orchestrating the computation and output of various metrics. It also allows me to selectively output desired metrics by commenting out unwanted print statements, ensuring a focused analysis.

**Testing**

To ensure the functionality and accuracy of the implementation, I designed several tests. (*Each test was computed by hand using a scientific calculator and cross referenced with the functions from rust to ensure validity***).**

1. **Reading the Graph:** The ‘test\_read\_graph’ function reads a test dataset into a graph, asserting that it contains 7 nodes. This verifies that the graph is read correctly from the dataset, ensuring its accurate representation.
2. **Jaccard Similarity:** The ‘test\_compute\_jaccard\_similarity’ function calculates the similarity between pairs of nodes in the test dataset, asserting that the similarity between nodes 3 and 2 is 0.25. This verifies the accuracy of the similarity computation by comparing it to a known value.
3. **Mean and Max Similarity:** The ‘test\_find\_mean\_max\_similarity’ function computes the mean and maximum similarity scores for the test dataset, asserting values of 0.075 and 0.25, respectively. This verifies that the statistical metrics are computed accurately and align with manually calculated values.
4. **BFS Shortest Path:** The ‘test\_bfs\_shortest\_path’ function computes the shortest path from a source vertex to a destination vertex in the test dataset, asserting a distance of 1. This ensures that the BFS algorithm accurately computes the shortest path.
5. **Similarity Percentages:** The ‘test\_compute\_similarity\_percentage’ function calculates the percentage of pairs with similarity scores above various thresholds, asserting a length of 10, covering thresholds from 0.1 to 1.0. It also checks specific values at 0.1 and 0.2, ensuring that the percentage calculations align with expected values.

**This is a screenshot of my test running successfully:**

*A screen shot of a computer

Description automatically generated*

**Output Analysis**

For pairs with shortest paths of 2, the similarity scores range from 0 to 1, reflecting varying degrees of similarity between researchers. For example, the Jaccard similarity between 20940 and 18215 is 0.0625, between 20552 and 25086 is 0.2727, and between 20517 and 8630 is 0.3333. These scores indicate different levels of similarity between authors’ co-authorship networks, from minimal connection to moderate overlap. This illustrates the diverse nature of the GR-QC research network, showing that while some pairs have closely related co-author networks, others have more distinct collaboration patterns.

**THIS IS A SNIPET OF MY CODE (SINCE IT IS TOO LONG TO PRINT THE WHOLE THING HERE):**

*Jaccard Similarity between 6033 and 17089: 0.0909*

*Jaccard Similarity between 25491 and 22826: 0.1667*

*Jaccard Similarity between 676 and 25272: 0.0385*

*Jaccard Similarity between 18487 and 9360: 0.0278*

*Jaccard Similarity between 24114 and 18676: 0.1111*

*Jaccard Similarity between 24592 and 13328: 0.1000*

*Jaccard Similarity between 6830 and 4446: 0.0208*

*Jaccard Similarity between 9736 and 11241: 0.0192*

*Jaccard Similarity between 6156 and 16485: 0.2000*

*Jaccard Similarity between 12733 and 4139: 0.1667*

*Jaccard Similarity between 17330 and 23387: 0.1176*

In terms of statistical analysis, the mean Jaccard similarity is 0.100, indicating an overall low similarity across the dataset. This suggests that, on average, there is limited overlap between the co-authorship networks of authors in the GR-QC field. The maximum Jaccard similarity is 1.000, with several pairs achieving this score, suggesting identical co-authorship networks. This indicates that certain pairs of authors have fully overlapping research networks, potentially representing well-established collaborations, or tightly knit research groups.

**THIS IS A SNIPET OF MY CODE (SINCE IT IS TOO LONG TO PRINT THE WHOLE THING HERE):**

*Mean Jaccard similarity: 0.100*

*Max Jaccard similarity: 1.000*

*Pairs with the highest similarity:*

*Pair: (25601, 21316), Similarity: 1.000*

*Pair: (21438, 10526), Similarity: 1.000*

*Pair: (17797, 17401), Similarity: 1.000*

*Pair: (3196, 23226), Similarity: 1.000*

*Pair: (6631, 114), Similarity: 1.000*

*Pair: (17924, 6702), Similarity: 1.000*

*Pair: (2664, 2081), Similarity: 1.000*

*Pair: (939, 14037), Similarity: 1.000*

*Pair: (22149, 9387), Similarity: 1.000*

The distribution of similarity scores above various thresholds shows a decreasing trend, with 28.45% of pairs above 0.1, 9.47% above 0.2, and 5.36% above 0.3. This reflects the generally low similarity across the network, highlighting areas for potential improvement. It shows that while there are some strong collaborations, many pairs of authors have limited overlap in their co-author networks, indicating opportunities for strengthening collaboration in the GR-QC field.

**THIS IS THE CODE:**

*Percentage of pairs with Jaccard similarity above 0.1: 28.45309068089112%*

*Percentage of pairs with Jaccard similarity above 0.2: 9.465014119861939%*

*Percentage of pairs with Jaccard similarity above 0.3: 5.362409789770944%*

*Percentage of pairs with Jaccard similarity above 0.4: 2.706306871666144%*

*Percentage of pairs with Jaccard similarity above 0.5: 1.0778161280200815%*

*Percentage of pairs with Jaccard similarity above 0.6: 0.8660181989331659%*

*Percentage of pairs with Jaccard similarity above 0.7: 0.7138374647003451%*

*Percentage of pairs with Jaccard similarity above 0.8: 0.671477878882962%*

*Percentage of pairs with Jaccard similarity above 0.9: 0.6557891433950424%*

*Percentage of pairs with Jaccard similarity above 1.0: 0%*

**Conclusion**

The analysis reveals varying collaboration strengths across the Arxiv GR-QC network. While many pairs exhibit low similarity scores, indicating limited collaboration beyond immediate co-authors, several pairs achieve high or even identical scores, suggesting opportunities for stronger co-authorship networks.

**Future Directions**

To further strengthen collaboration in the GR-QC network, several steps can be taken. First, encouraging authors to leverage their mutual contacts through conferences, workshops, and online forums dedicated to the GR-QC field can foster new collaborations and expand co-authorship networks. Second, monitoring the network over time can help analyze changes in similarity scores and identify emerging trends in co-authorship patterns. This will allow us to gauge the success of initiatives aimed at fostering collaboration and identify areas where further support is needed. Lastly, extending the analysis to other Arxiv categories can help explore how collaboration patterns compare across different research domains. This can reveal whether the trends observed in the GR-QC network are specific to this field or reflect broader patterns in academic research, potentially offering insights into how to strengthen collaborations across various domains.