

Report of Email Network Dataset

In this analysis, I explored the properties of an email network dataset obtained from the email-Eu-core dataset. The goal was to understand the structure of the network and gather insights regarding the average distance between pairs of vertices, average outdegree, and shortest paths.

Algorithms Used

Breadth-First Search (BFS): BFS is a graph traversal algorithm that explores all the vertices of a graph in breadth-first order, starting from a given source vertex. It computes the shortest path distances from the source vertex to all other vertices in the graph.

Data Preprocessing

The email network dataset was read from the "email-Eu-core.txt" file. The file contained a list of edges representing email communication between individuals. I processed each line in the dataset, extracting the source and target nodes of each edge. I created an adjacency list representation of the graph and calculated the number of nodes and edges in the network.

Analysis Results

Number of Nodes: The email network dataset consisted of 1005 nodes.

Number of Edges: The dataset contained a total of 25571 edges.

Average Outdegree: The average outdegree, which represents the average number of outgoing edges from each node, was calculated to be 25.46.

Average Distance of Vertices: To determine the average distance between pairs of vertices, we performed BFS from each node in the graph and calculated the shortest path distances. The average distance was found to be 3.691.

Discoveries

The email network dataset exhibited a relatively high average outdegree, indicating a dense network where individuals had a substantial number of email connections.

The average distance of 3.691 suggested that, on average, it took approximately 3.691 steps (or hops) to reach any other vertex in the network from a given vertex. This value provides insights into the overall connectivity and reachability within the email network.

Conclusion

In conclusion, the analysis of the email network dataset using BFS and other graph analysis techniques revealed valuable insights into the structure and characteristics of the network. The average outdegree and average distance metrics provided an understanding of the network's density and connectivity. These findings contribute to our knowledge of email communication patterns and network dynamics.

Dataset

<https://snap.stanford.edu/data/email-Eu-core.html>

Sources

<https://www.ericburden.work/blog/2022/12/14/advent-of-code-day-14/>

<https://doc.rust-lang.org/std/collections/struct.HashSet.html>