DS 210 Final Project
Tami Ajibade

I explored network analysis through a dataset representing an email communication network. The primary objective of my project was to apply the concept of 'six degrees of separation' within this network, analyzing the average number of steps required for any given node to connect with every other node in the data set. The project's crux was to delve into the fascinating realm of social network analysis by employing a dataset from a substantial European research institution made available through the SNAP database. The dataset consists of email interactions establishing a complex web of connections representative of real-world communication patterns.

To conduct this analysis, I implemented several functions in Rust, each serving a pivotal role in deciphering the network's properties.

Data Processing

- `read_file` Function: A utility to process the dataset file, which parses each line (excluding comments marked with '#') to construct a collection of tuples representing the network's edges.

Graph Construction

- `adjacency_list` Function: Transforms the edge list into a hashmap where each key is a node, and the corresponding value is a vector containing adjacent nodes (i.e., direct connections).

Network Analysis

- `bfs` Function (Breadth-First Search): Accepts a starting node and the adjacency list, returning a hashmap that correlates each node with its distance from the start node.
- `bfs_all_nodes` Function: Utilizes the bfs function to create a hashmap where the key is the node, and the value is the average distance from this node to all other nodes in the network.
- `average_degrees_of_separation` Function: Leverages the bfs_all_nodes output to calculate the mean value of separation degrees across the entire network.

Upon executing the `bfs_all_nodes` function and analyzing its output, some intriguing patterns emerged. Several nodes exhibited an average distance of exactly one degree. Closer inspection of the dataset suggested these represent entities engaged in exclusive communication with a single other node, hinting at external email addresses corresponding to limited interaction within the network. Additionally, the presence of NaN values among the average distances pointed to nodes that exclusively communicated with themselves, indicating self-loops without further network integration.

The tool successfully provided a holistic view of the network's interconnectivity. By examining the distances and connectivity patterns, it could infer the degrees of separation within the network, adhering to the 'six degrees of separation' theory. The project underscored the significance of network analysis in understanding communication dynamics. The developed tool not only calculates fundamental network statistics but also provides a lens through which the nuances of social interactions can be examined quantitatively. Notably, the findings emphasized the varied nature of connectivity within the network, from isolated interactions to more integrated communication.

DS 210 Final Project
Tami Ajibade

```
    Finished dev [unoptimized + debuginfo] target(s) in 2.23s
     Running `target/debug/final_project`
Graph Summary:
Total nodes: 914
Total edges: 56507
Graph density: 0.0339
Average degrees of separation: 3.42715731985373
Node with the highest identifier: 1000
Node with the lowest identifier: 0
Number of self-loops: 313
Nodes with exactly one connection: [825, 961, 51, 593, 156, 474, 111, 172, 864, 790, 857, 44, 428, 492, 31, 129, 70
9, 616, 594, 310, 377, 7, 482, 869, 127, 371, 57, 914, 321, 928, 268, 36, 416, 642, 699, 405, 986, 141, 679, 904, 5
60, 894, 271, 946, 357, 600, 234, 105, 899, 686, 988, 839, 767, 445, 976, 6, 250, 369, 835, 546, 291, 183, 436, 641
, 343, 394, 12, 756, 556, 191, 84, 487, 777, 74, 221, 948, 165, 637, 962, 241, 100, 170, 21, 646, 896, 95, 470, 88,
 348, 688, 368, 458, 773, 82, 711, 806, 831, 128, 768, 817, 983, 159, 824, 249, 813, 734, 322, 735, 866, 73, 521, 5
26, 854, 365, 453, 226, 281, 789, 39, 807, 876, 574, 753, 624, 728, 582, 2, 329, 706, 270, 804, 507, 936, 18, 293,
437, 511, 153, 295, 319, 411, 300, 78, 382, 826, 909, 266, 331, 508, 168, 370, 514, 963, 448, 419, 610, 727, 297, 4
07, 353, 532, 759, 987, 29, 361, 494, 375, 643, 979, 228, 774, 513, 625, 955, 475, 265, 490, 244, 989, 125, 90, 550
, 205, 381, 208, 354, 960, 667, 144, 867, 829, 970, 61, 443, 754, 926, 719, 958, 763, 204, 284, 260, 210, 905, 793,
 359, 858, 43, 696, 520, 570, 971, 311, 651, 384, 583, 542, 533, 378, 218, 449, 568, 246, 396, 576, 608, 927, 708,
239, 552, 174, 554, 796, 993, 787, 263, 820, 964, 655, 324, 69, 164, 374, 680, 429, 558, 471, 587, 423, 345, 166, 6
48, 892, 179, 286, 103, 341, 722, 484, 216, 145, 80, 785, 229, 591, 639, 420, 367, 121, 687, 142, 995, 750, 150, 35
1, 916, 613, 67, 214, 496, 548, 339, 412, 871, 189, 540, 194]
The graph is not a complete graph.
crc-dot1x-nat-10-239-25-113:trial tamiajibade$
```

The output generated by the tool upon execution is a comprehensive statistical summary of the email network's properties, with key metrics highlighted for a quick overview:

- ➔ Total nodes: This represents the total number of unique entities (individual email addresses) within the network.
- ➔ Total edges: The number of direct email interactions present in the dataset. Each edge implies a communication link between two nodes.
- ➔ Graph density: A measure reflecting how complete the graph is. A higher density indicates a more interconnected network, while a lower density suggests sparser connections.
- ➔ Average degrees of separation: An average indicating how many steps are needed to connect any two nodes in the network. The closer this number is to six, the more the dataset adheres to the six degrees of separation theory.
- ➔ Node with the highest identifier: The largest numerical identifier found among all nodes, which can be helpful in assessing the range of the dataset.
- ➔ Node with the lowest identifier: The smallest numerical identifier, which is often zero in zero-indexed datasets.
- ➔ Number of self-loops: A count of nodes that have sent emails to themselves, which could be indicative of test accounts or administrative actions.

- ➔ Nodes with exactly one connection: These nodes represent email addresses with only a single recorded interaction, possibly indicating limited participation in the network or external contacts.

The output also specifies whether the graph represents a complete network. In a complete graph, every pair of distinct vertices is connected by a unique edge. If the network is not complete, it implies that not all nodes are directly connected to each other.