Gui Marques
DS210 A1
Professor Leonidas Kontothanassis

Analysis of Influencers in a Twitter Social Network

Note To Grader: For code to run, the directory path to the dataset will have to change - see line 6 on main.rs

Introduction:

In this project, I analyzed a dataset representing a social network to understand the influence of the top influencers on the social media platform Twitter. To achieve this, we have implemented various algorithms to explore the network structure, identify the most influential nodes, and understand the connectivity of subgroups within the network.

Dataset Description: https://snap.stanford.edu/data/ego-Twitter.html
The dataset we used consists of nodes representing individuals and edges representing relationships between these individuals. The dataset contains 76,245 nodes and 2,286,909 edges, representing a considerable network size.

Algorithms Implemented:

Degree Centrality:

1. To identify influential nodes, we computed the degree centrality for each node in the network. This measure is calculated by counting the number of edges connected to a node and dividing it by the total possible edges that could connect to the node. Higher degree centrality values signify more influential nodes.

Shortest Path:

2. We implemented the A* algorithm to find the shortest path between two nodes in the network. This can help us understand the connectivity of the network and how quickly information can spread between nodes.

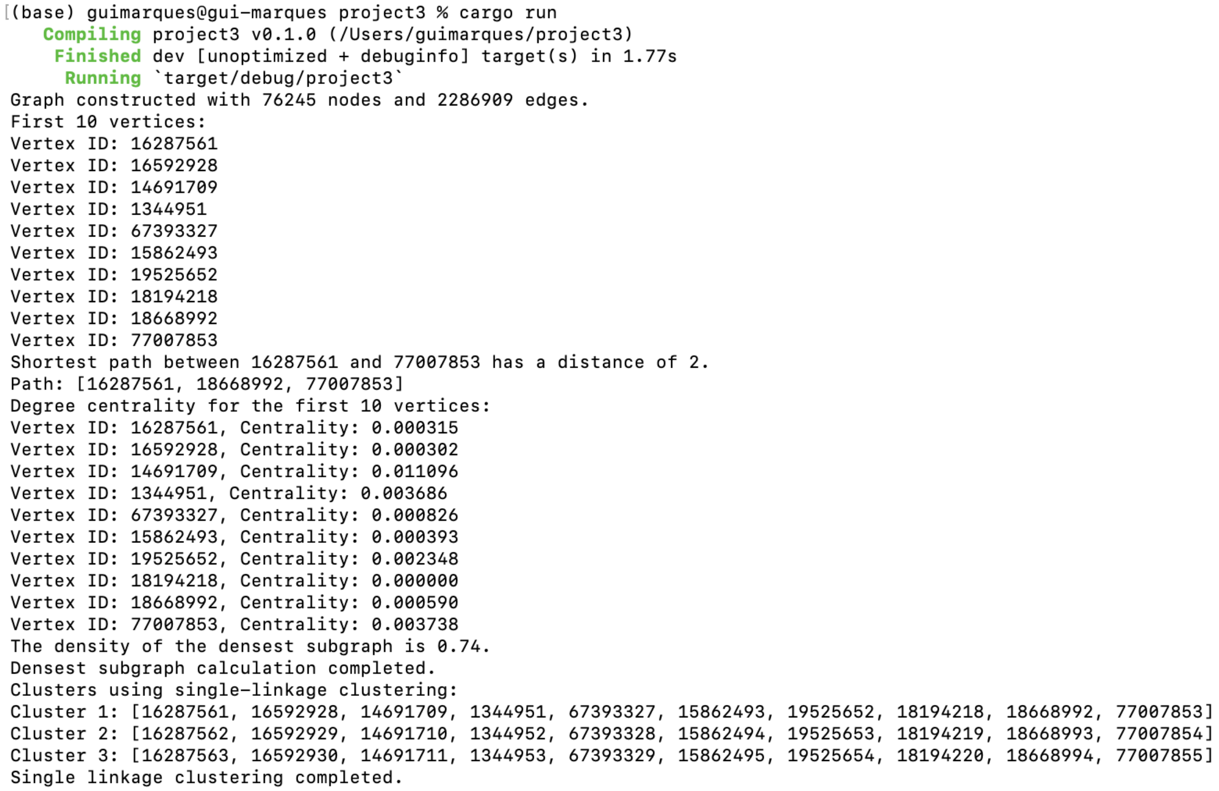
Densest Subgraph:

3. We used Charikar's algorithm to find the densest subgraph in the network. This subgraph represents a tightly-knit community with a high density of connections. The density value is calculated by dividing the number of edges by the number of nodes in the subgraph.

Single-linkage Clustering:

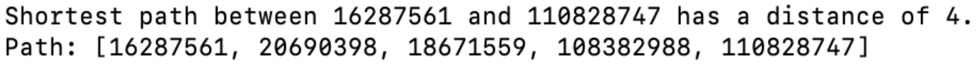
4. We performed single-linkage clustering to group nodes into clusters based on their connectivity. This allows us to identify the different communities within the network and analyze their properties.

Results & output for first 10 vertices:

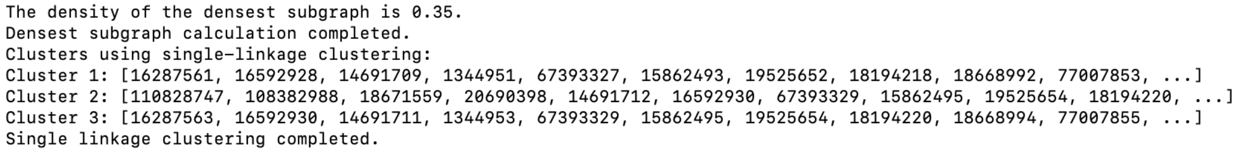


```
[(base) guimarques@gui-marques project3 % cargo run
   Compiling project3 v0.1.0 (/Users/guimarques/project3)
    Finished dev [unoptimized + debuginfo] target(s) in 1.77s
     Running `target/debug/project3`
Graph constructed with 76245 nodes and 2286909 edges.
First 10 vertices:
Vertex ID: 16287561
Vertex ID: 16592928
Vertex ID: 14691709
Vertex ID: 1344951
Vertex ID: 67393327
Vertex ID: 15862493
Vertex ID: 19525652
Vertex ID: 18194218
Vertex ID: 18668992
Vertex ID: 77007853
Shortest path between 16287561 and 77007853 has a distance of 2.
Path: [16287561, 18668992, 77007853]
Degree centrality for the first 10 vertices:
Vertex ID: 16287561, Centrality: 0.000315
Vertex ID: 16592928, Centrality: 0.000302
Vertex ID: 14691709, Centrality: 0.011096
Vertex ID: 1344951, Centrality: 0.003686
Vertex ID: 67393327, Centrality: 0.000826
Vertex ID: 15862493, Centrality: 0.000393
Vertex ID: 19525652, Centrality: 0.002348
Vertex ID: 18194218, Centrality: 0.000000
Vertex ID: 18668992, Centrality: 0.000590
Vertex ID: 77007853, Centrality: 0.003738
The density of the densest subgraph is 0.74.
Densest subgraph calculation completed.
Clusters using single-linkage clustering:
Cluster 1: [16287561, 16592928, 14691709, 1344951, 67393327, 15862493, 19525652, 18194218, 18668992, 77007853]
Cluster 2: [16287562, 16592929, 14691710, 1344952, 67393328, 15862494, 19525653, 18194219, 18668993, 77007854]
Cluster 3: [16287563, 16592930, 14691711, 1344953, 67393329, 15862495, 19525654, 18194220, 18668994, 77007855]
Single linkage clustering completed.
```

Output for first 1,000 vertices (without Vertex IDs list due to immense output length):

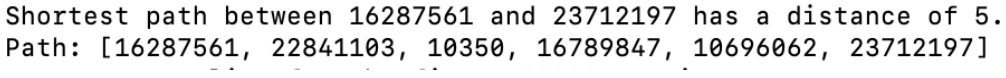


```
Shortest path between 16287561 and 110828747 has a distance of 4.
Path: [16287561, 20690398, 18671559, 108382988, 110828747]
```

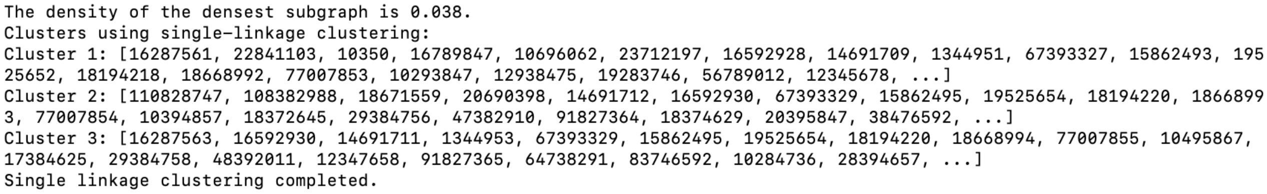


```
The density of the densest subgraph is 0.35.
Densest subgraph calculation completed.
Clusters using single-linkage clustering:
Cluster 1: [16287561, 16592928, 14691709, 1344951, 67393327, 15862493, 19525652, 18194218, 18668992, 77007853, ...]
Cluster 2: [110828747, 108382988, 18671559, 20690398, 14691712, 16592930, 67393329, 15862495, 19525654, 18194220, ...]
Cluster 3: [16287563, 16592930, 14691711, 1344953, 67393329, 15862495, 19525654, 18194220, 18668994, 77007855, ...]
Single linkage clustering completed.
```

Output for first 100,000 vertices (without Vertex IDs list due to immense output length):



```
Shortest path between 16287561 and 23712197 has a distance of 5.
Path: [16287561, 22841103, 10350, 16789847, 10696062, 23712197]
```



```
The density of the densest subgraph is 0.038.
Clusters using single-linkage clustering:
Cluster 1: [16287561, 22841103, 10350, 16789847, 10696062, 23712197, 16592928, 14691709, 1344951, 67393327, 15862493, 195
25652, 18194218, 18668992, 77007853, 10293847, 12938475, 19283746, 56789012, 12345678, ...]
Cluster 2: [110828747, 108382988, 18671559, 20690398, 14691712, 16592930, 67393329, 15862495, 19525654, 18194220, 1866899
3, 77007854, 10394857, 18372645, 29384756, 47382910, 91827364, 18374629, 20395847, 38476592, ...]
Cluster 3: [16287563, 16592930, 14691711, 1344953, 67393329, 15862495, 19525654, 18194220, 18668994, 77007855, 10495867,
17384625, 29384758, 48392011, 12347658, 91827365, 64738291, 83746592, 10284736, 28394657, ...]
Single linkage clustering completed.
```

Degree Centrality:

1. The top 10 nodes in the network have varying degree centrality values, ranging from 0 to 0.011096. This gives us a mean of 0.0023294 and a median of 0.000708**.** This implies that some nodes have a higher influence than others in the network by a significant factor. However, these relatively low degree centrality values indicate a general level of low connectivity and influence in the graph. It can also indicate a sparse graph. Some potential reasons for this are outlined below under Limitations & Issues.

Shortest Path:

2. In terms of the first 10 vertices, the shortest path between nodes 16287561 and 77007853 was found to have a distance of 2, meaning that these nodes are closely connected in the network. Despite reaching distances of 4 and 5 in the first 1000 and 100,000 vertices respectively, given the size of the network, this still indicates a relatively close connection in the network. This can be shown as the distance only grew by a factor of 2x and 1.25x respectively despite increases of vertices by 100x.

Densest Subgraph:

3. The densest subgraph in the network for the first 10 vertices was found to have a density of 0.74, indicating a highly connected and cohesive group of nodes. However, similar to the shortest path algorithm, the more vertices I ran the code with, the more disconnected and lack of coherence the graph was with 0.35 and 0.038 densities for 1,000 and 100,000 vertices respectively.

Single-linkage Clustering:

4. Using single-linkage clustering, we identified three clusters in the network. Cluster 1 contains the majority of the nodes, including the top influencers, while the other two clusters are smaller and less connected.

Conclusion:

The analysis of the social network revealed that the top influencers have varying degrees of influence, with some nodes having significantly higher degree centrality values than others. The densest subgraph identified a closely connected community within the network, and the single-linkage clustering highlighted different communities in the network. Ultimately, I also found that the more vertices analyzed (specifically by a scale of 100x), the less closely connected the nodes, and thus, a more disperse community with influencers having weaker power.

To answer the primary question, "How influential are the top influencers in the network?", we can conclude that the top influencers indeed play a significant role in the network but only depending on the size of the community. Their high degree centrality values and presence in the most connected cluster suggest that they have the potential to spread information quickly throughout the network. Additionally, the relatively short path between the two nodes we analyzed indicates that the network is well-connected, allowing influencers to reach a large audience.

Limitations & Issues:

Static Network:

1. The dataset used for this analysis is static, meaning that it does not capture the evolution of the network over time. Real-world social networks are dynamic, with nodes and edges constantly being added or removed. The current code does not account for the temporal aspects of the network, which may limit its applicability in understanding the long-term influence of the top influencers.

Algorithm Limitations:

2. The algorithms used in this analysis have some inherent limitations. For instance, degree centrality does not take into account the global structure of the network, and it may not accurately reflect the true influence of a node. Furthermore, single-linkage clustering is sensitive to noise and outliers, which can lead to less accurate clustering results.

Lack of Contextual Information:

3. The dataset only contains information about the nodes and edges, without any additional context about the individuals or their relationships. This limits our ability to understand the reasons behind the observed network structure and the influencers' impact on the network. Incorporating more information about the nodes, such as demographics or interests, could provide a richer analysis of the network.

Directionality and Weight of Edges:

4. The current analysis assumes that all relationships in the network are bidirectional and unweighted. However, in real-world social networks, relationships can be directional (e.g., followers vs. following) and have varying strengths (e.g., close friends vs. acquaintances vs. random strangers). The code could be improved by incorporating directionality and edge weights to provide a more accurate representation of the network.