DA3018 Project

Adriel Alander

June 1, 2020

1 Introduction

Currently the graph include many incorrect edges. Due to the large file, a visual representation is not appropriate, however it is possible to investigate the properties of the graph.

The properties investigated in this project include the following:

- The node degree distribution.
- The number of components.
- The component size distribution.

When analysing this, a java program was constructed. This will generate a simplified version of the graph, as well as two `txt`-file consisting the degree distribution as well as the component distribution. These are included in the GitHub repository that can be found here.

2 Algorithms and Methods

2.1 Reading File

For reading each file, `java.io.BufferedReader` was used, as it appeared to be the quickest methods for reading a file. Each node got assigned an `Integer` using a hash map. It's chosen to work with this instead of the strings, as they use less memory. Using the translated vertices, a new `txt`-file is created, only consisting of the node and edge information.

2.2 Creating Graph

For storing each node and it's neighbors, an adjacency list was used.

2.3 Degree Distribution

Due to the adjacency list, the degree distribution is already calculated upon creation of graph. This information was printed and stored in a file, and later used to print a histogram.

2.4 Component Distribution

In order to find the number of components as well as how many vertices are included in each component, *breadth-first search* was used. As it was necessary to examine each node and edge, this method appeared to be the best option available, with a complexity of $O(|V||E|)$, where $|V|$ is the number of vertices, and $|E|$ is the number of edges.

3 Results

3.1 Degree Distribution

The total amount of vertices is 11393435. Figure 1 shows the degree distribution.

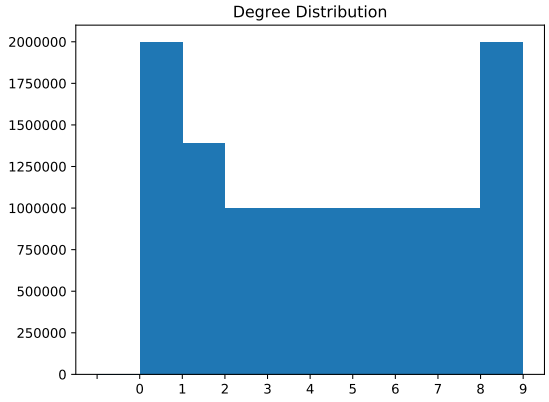

Figure 1: The degree distribution of all nodes in graph. The x-axis represent the degree, and the y-axis the number of nodes with this degree.

3.2 Component Distribution

The total amount of components is 496481. These appear to differ in sizes ranging from 2 vertices to 7282226. This was failed at representing it in any

visual matter.

4 Conclusions

Despite this project, I frankly have little to no knowledge of this subject. I am not a biologist, therefore I am unable to draw any conclusions. Hopefully the data will help, though I cannot guarantee it is correct, as I am not very familiar with the subject and am unable to confirm with any sort of reference.