Introduction to Science of Networks Project Report

Bhishmaraj Selvamani Roll No:1410110099

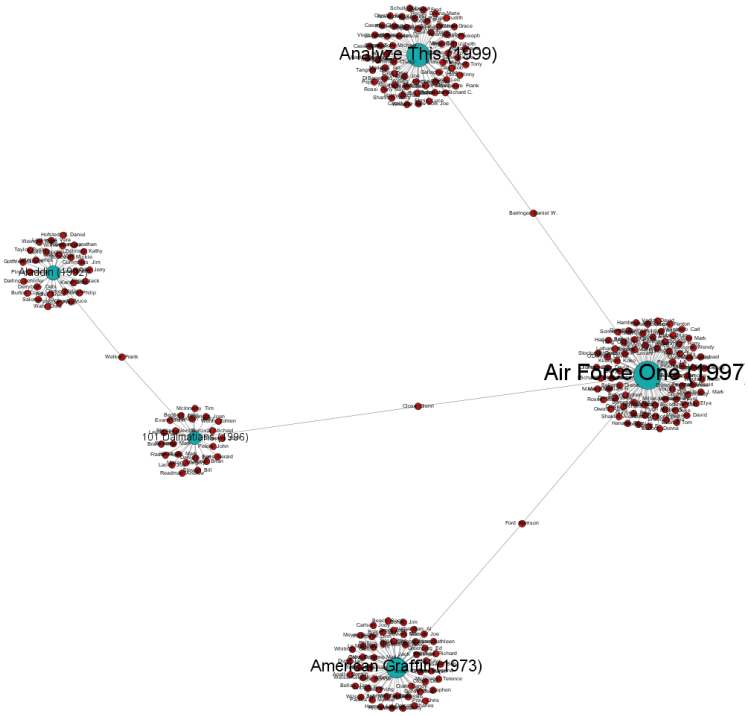
Analysis of the co-stardom network and the six degrees of Kevin Bacon

1. Introduction

The co-stardom network shows the collaboration network between film actors. It can be represented as undirected bipartite graph where the 2 disjoint set of nodes are actors and movies, therefore there exists an edge between an actor and a movie if he had starred in it.

There are also other ways of representing the graph such as connecting 2 actors if they had starred in the same movie but as the number of movies and actors grow it becomes very hard to analyze such a big graph, so I decided to analyze it as a bipartite graph.

As we can see it is a bipartite graph in which movies are colored blue and actors are colored red. There exists a node common between two movies if a particular actor acted in those 2 movies.



A tiny part of the actor-movie bipartite graph

I have also found the frequency of Bacon numbers in each of these datasets. Bacon Number is defined as the number of degrees of separation of a particular actor from Kevin Bacon. For example, Kevin bacon has a Bacon number of 0, all other actors who have co-acted with him have a Bacon number of 1. All other actors have Bacon number 1 greater than the previous actor how has been linked from Kevin Bacon.

I have found the average path length from Kevin Bacon to all other actors for various datasets.

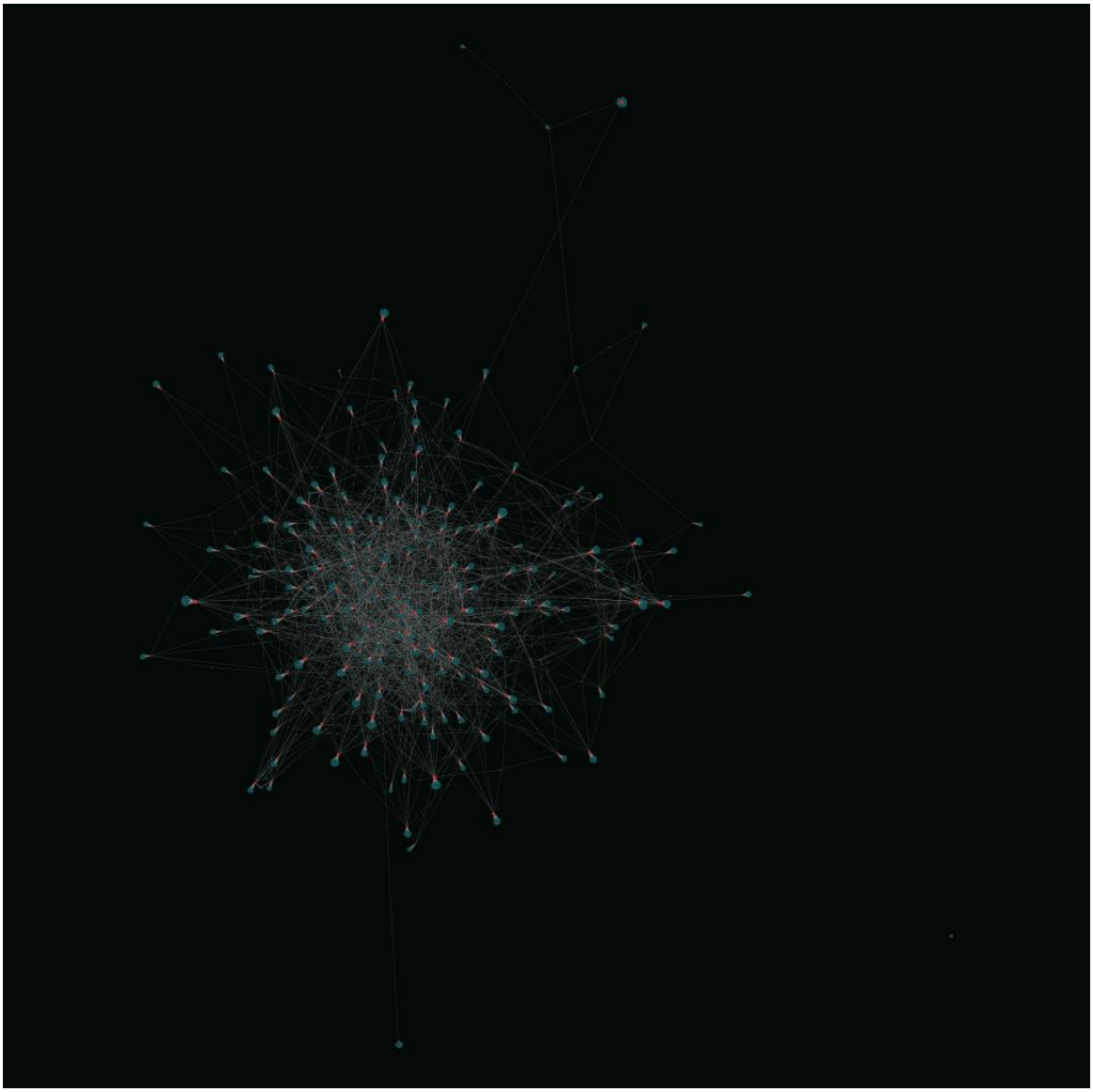
The raw data from IMDb contained the movie and the actors acted in it. So I created a java program to parse the raw data and convert it into 2 csv files. One csv file contains the following description of the nodes a unique ID, label, attribute an attribute describes weather a node is a movie or not. The other csv file contains the list of undirected edges between movies and actors.

Data Visualization

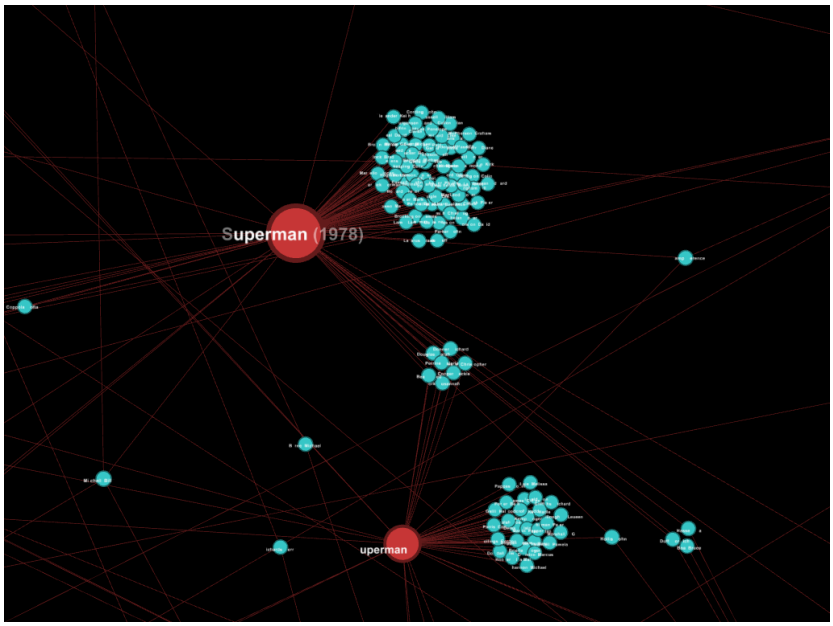
The following datasets were used in my analysis:

FILE	DESCRIPTION	MOVIES	ACTORS	EDGES
cast.00-06.txt	movies release since 2000	52195	348497	606531
cast.G.txt	movies rated G by MPAA	1288	21177	28485
cast.mpaa.txt	movies rated by MPAA	21861	280624	627061
cast.rated.txt	popular movies	4527	122406	217603
movies-top-grossing.txt	Top grossing movies	187	8264	9920

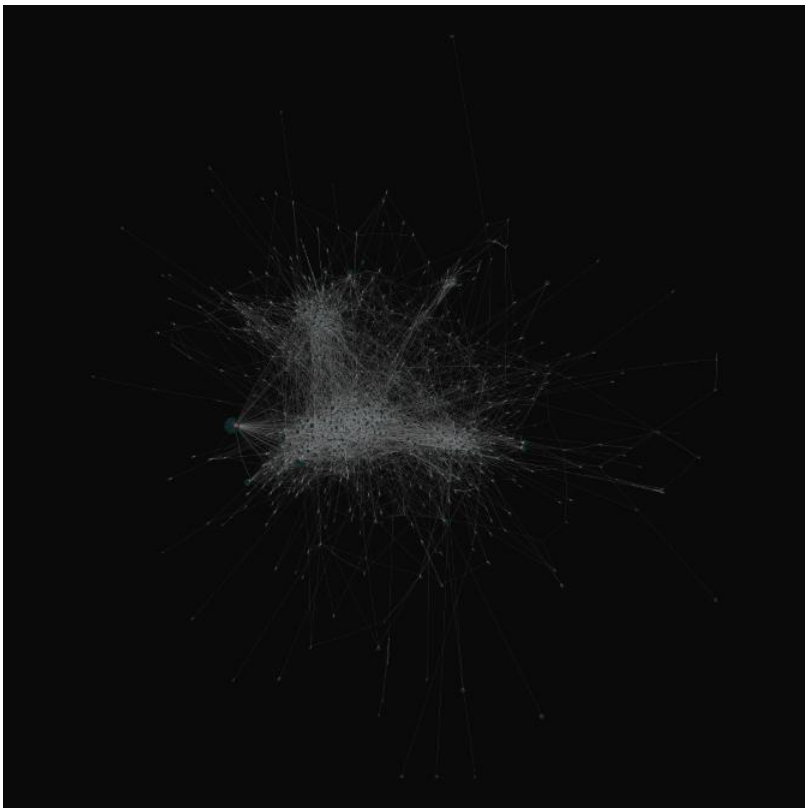
The data is taken from the Internet Movie Database*.*



Visualization of one of the Actor – Movie networks (movies-top-grossing.txt) using Gephi



I used Gephi to visualize my graph network and found some interesting findings such as the one above where we can see 2 movies sharing many actors and these 2 movies happen to be sequels.



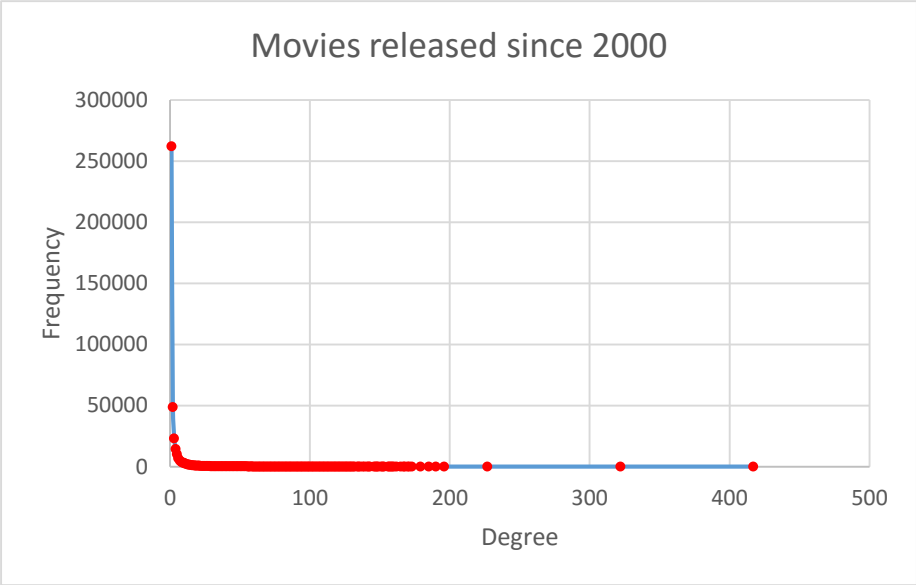
Visualization of an actor-movie network

Nodes: 22465

Edges: 28485

Analysis

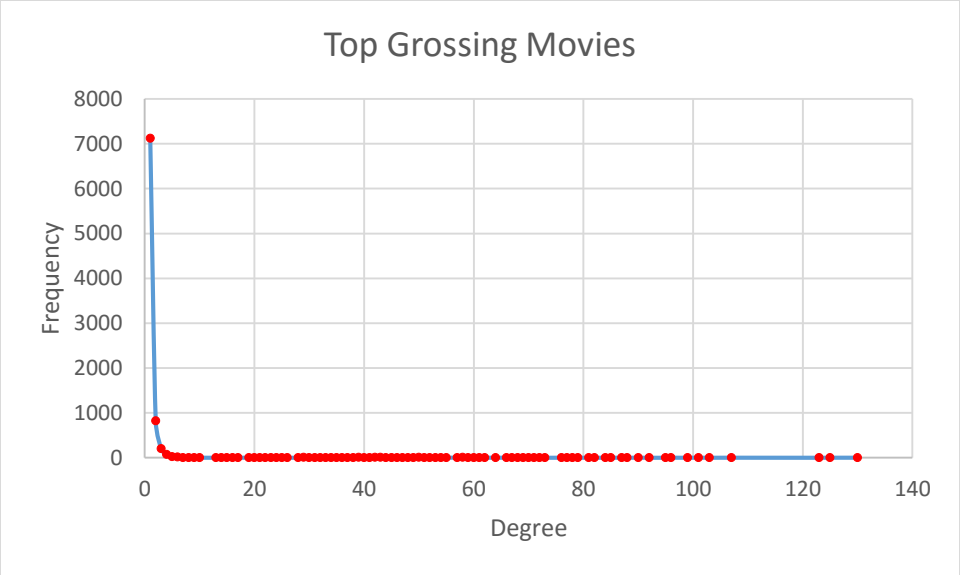
Degree Distributions



Nodes: 400692

Edges: 606531

Average degree: 3.027

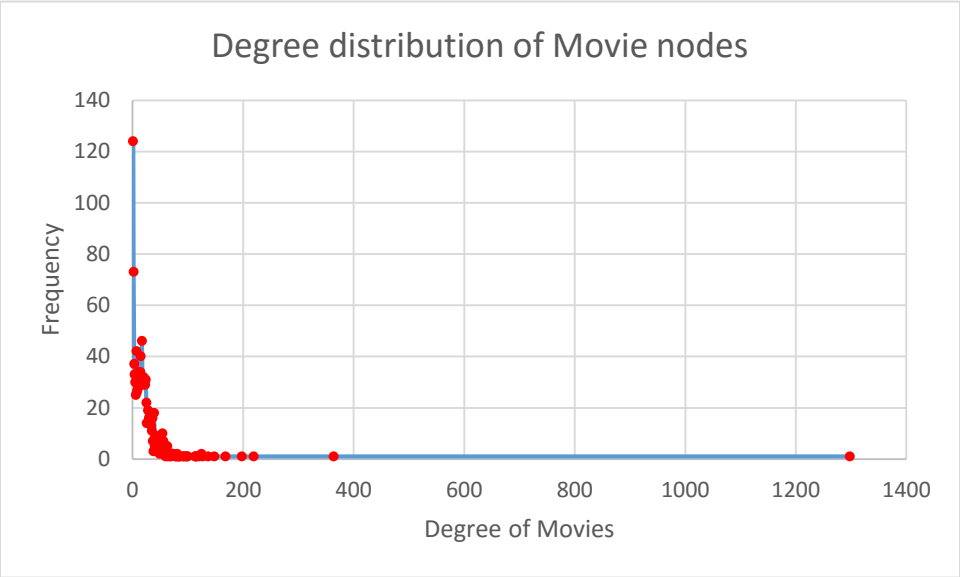


Nodes: 8451

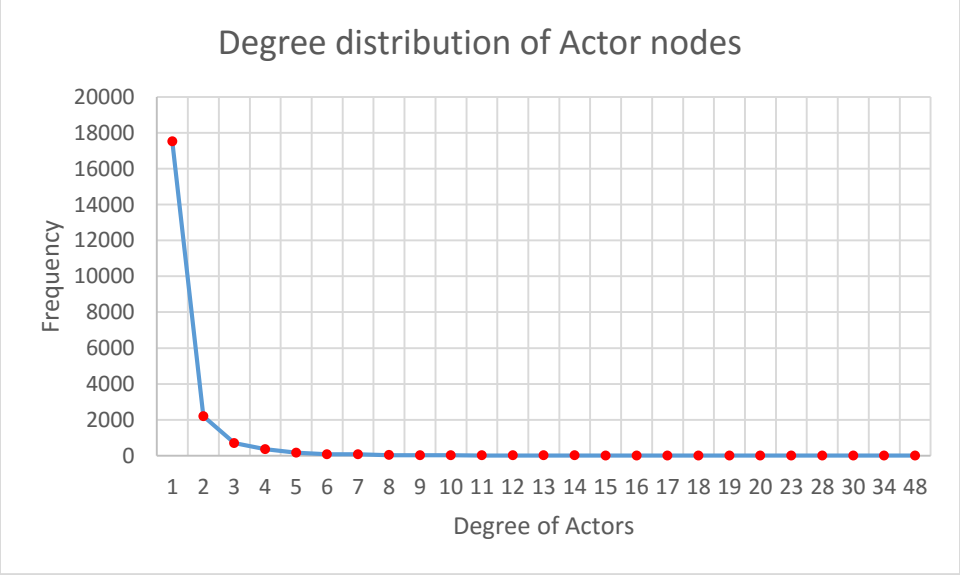
Edges: 9920

Average degree: 2.347

Due to the bipartite nature of the graph its more informative to plot 2 different degree distributions, one for the actor and another for movie.



Number of movies:1288



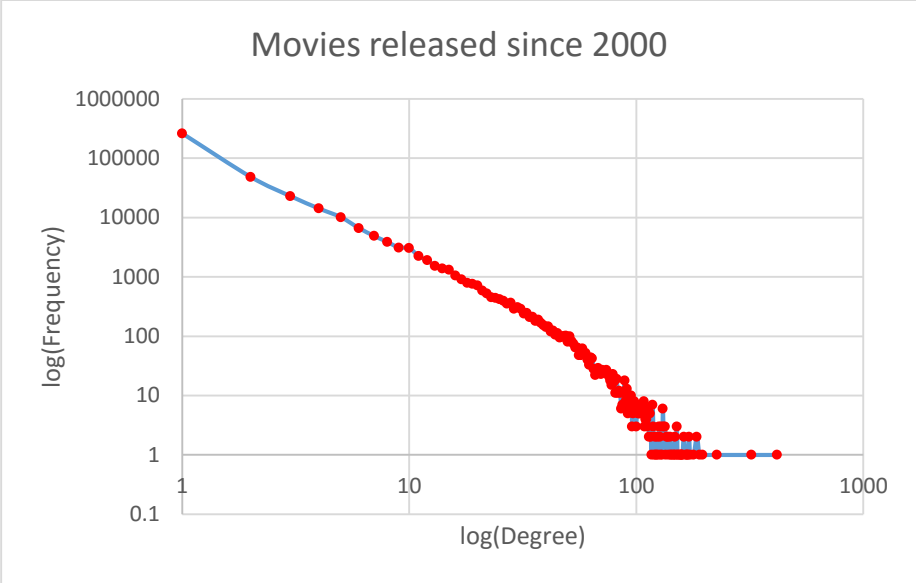
Actors: 21177

From the above graphs degree distributions follow a power law.

A function f(x) is known as a power law If it can be written in the form

$$f(x) = ax^{-k}$$

One aspect of a power law is that it follows scale invariance so scaling the power-law equation by a constant c it simply multiplies the function by a constant factor of c^{-k}. This property makes the power law to show a linear relationship when logarithm is taken on both the sides of the equation.

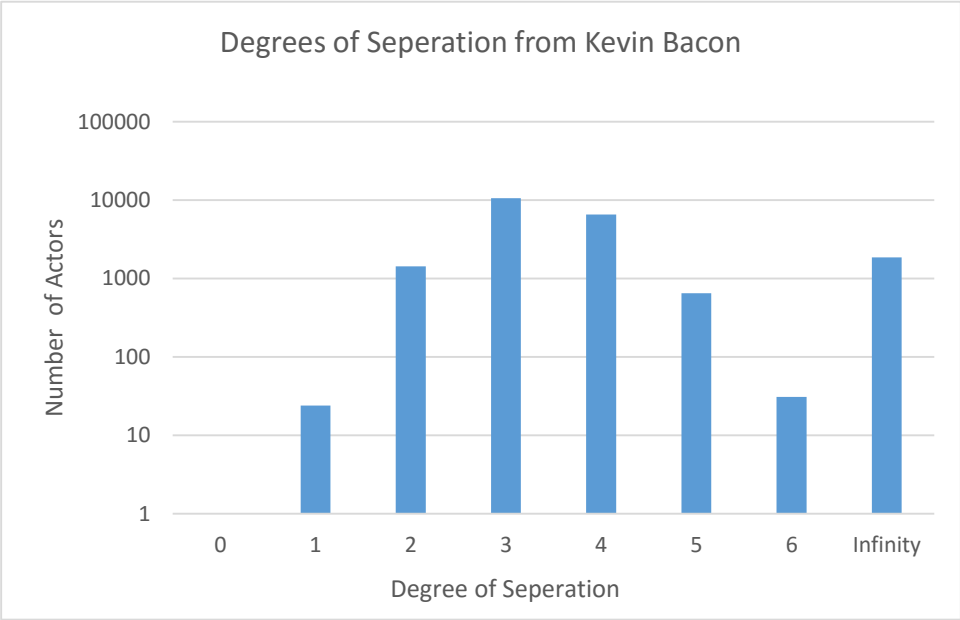


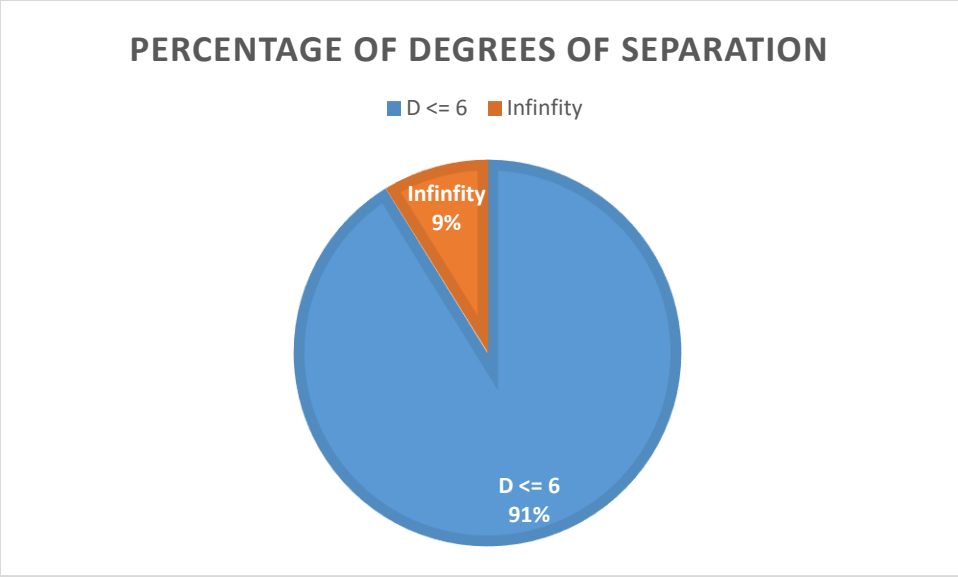
From the above log-log graph we can see the linear relationship of the frequency with the degree.

Distribution of Degrees of Separation

Keeping Kevin Bacon as the center of the graph, I wrote a program that ran a **Breath First Search** to all other reachable nodes and found the frequency of the path lengths. Breath First Search is used because it can be seen that **more than 90%** of the actors can be linked to Kevin Bacon in **less than 6 links.**

Total Actors: 21177

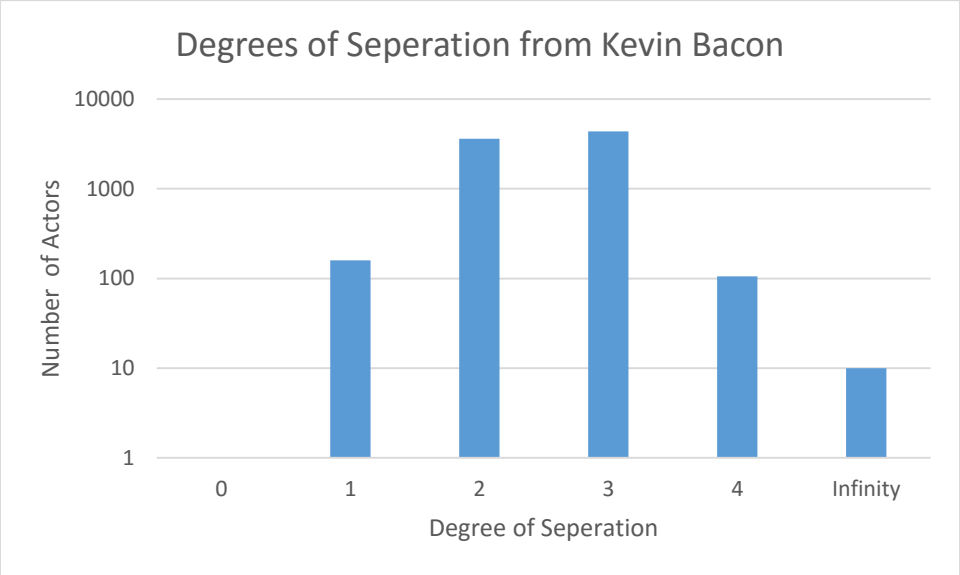




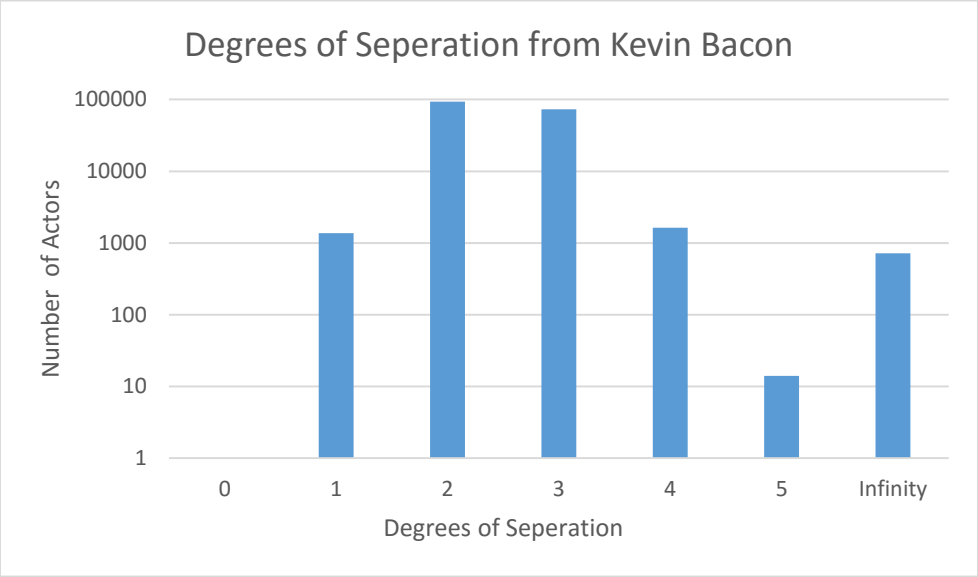
D: Degrees of Separation

Infinite degrees of separation mean that those actors aren't connected to Kevin Bacon.

Total actors: 8264



Total Actors: 170518



We can see that the six degrees of separation holds good for most of the datasets.

Conclusion

From the above analysis its quite clear even though our networks are really big in terms of size they are also small in terms of connectivity. This is the main principle of the small world phenomenon. It demonstrates the power of weak ties which makes these networks highly connected.

Bibliography

http://www.imdb.com/interfaces

https://en.wikipedia.org/wiki/Six_degrees_of_separation

http://introcs.cs.princeton.edu/java/45graph/

http://oracleofbacon.org/how.php

https://www.cs.cornell.edu/home/kleinber/swn.pdf

https://en.wikipedia.org/wiki/Power_law