INTROSPECTING DIFFERENT COMMERCIAL MOVIE RECOMMENDATION PLATFORM

SOCIAL COMPUTING TERM PROJECT REPORT

BY

TUSHAR SINHA (15IE10038)
ANUBHAV SHUKLA (15IM30002)

Group number:- SC13

UNDER THE SUPERVISION OF

PROF. SAPTARSHI GHOSH

MR. ABHISEK DASH

DEPARTMENT OF INDUSTRIAL AND SYSTEMS ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Introduction

What have we worked? We developed crawlers for NETFLIX and AMAZON Prime on the basis of previous developed crawlers for youtube. Further the crawler crawled starting from a random node(Movie, TVshow, etc) and stored the data such as name of movies, genre, production houses, etc. The crawler was crawled for ‘x’ number of times with one week gaps between 2 consecutive crawls. Next a directed graph was created with shows as the nodes and edges pointing from node to its recommended list of shows. A data for Netflix originals and Amazon Prime was extracted from various sites, this data was used to differentiate between Netflix originals or Amazon Prime Originals and other studios shows in the crawled data. Using NetworkX some analyses were done to check if Netflix and Amazon Prime shows biases towards their originals in their recommendation system.

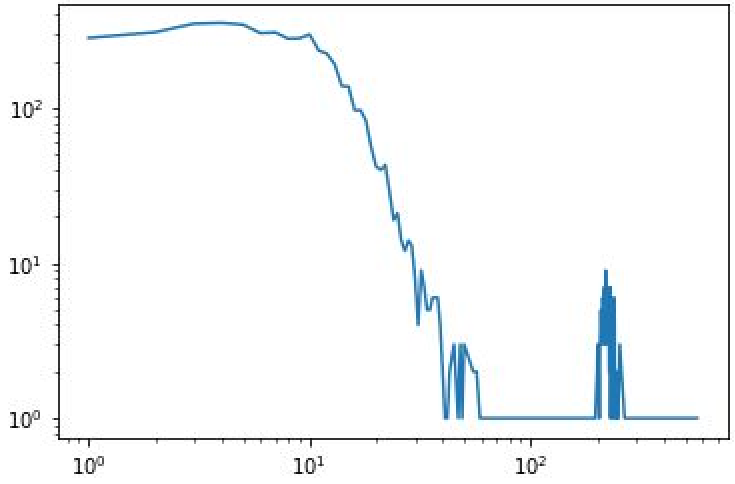
Further we what the are the different analysis performed, the theory behind them, the results and the conclusions based on these results. For Netflix we have crawled data for 3 different weeks in logged out mode. For Amazon Prime we have crawled data only once since it’s recommendation was same for every different week.

Analysis

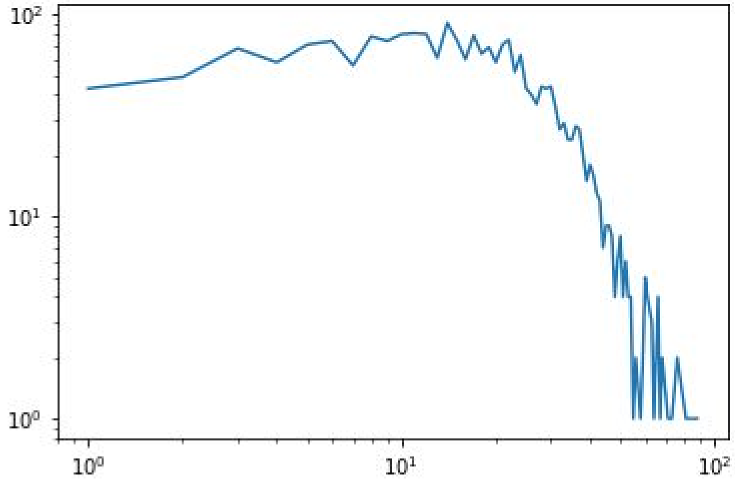
1. Degree Distribution Log-Log Plot

After creating the graphs using networkX, we got the indegree of all the nodes of that graph. Once, then for each indegree value k, we found out the number of nodes having degree k. Then using matplotlib plot the InDegree Distribution Log-Log Plot was done, where X-Axis is Indegree (sorted in decreasing order) and Y-axis is the degree value.

1.1. NETFLIX

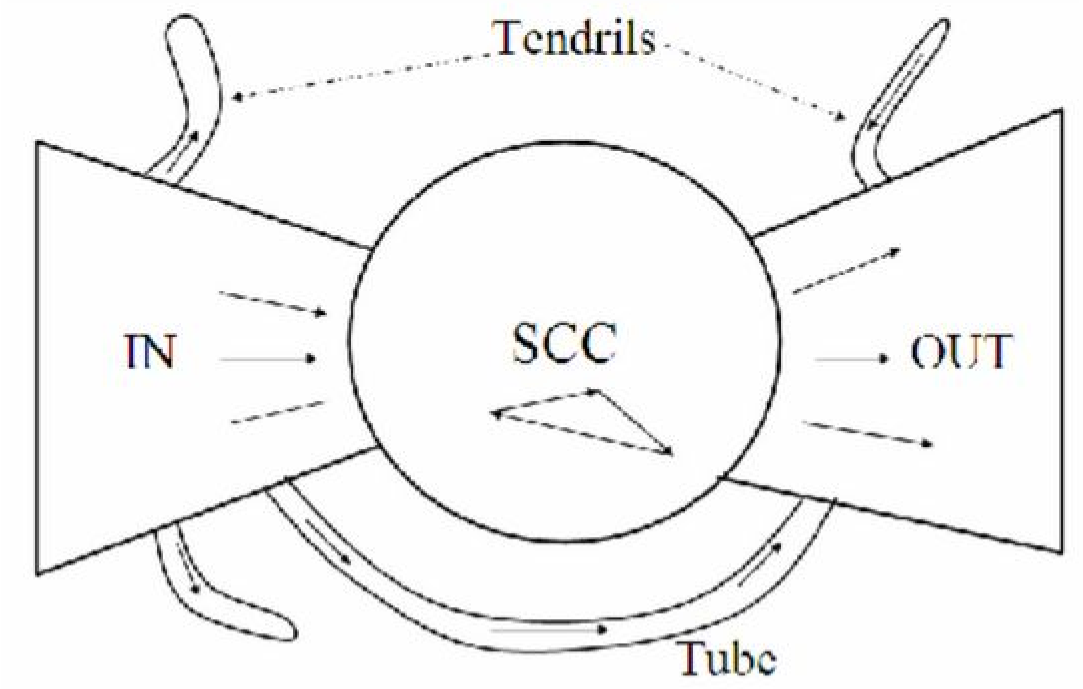


1.2. Amazon Prime



2. Bow-Tie Analysis

Any WebGraph is expected to follow the BowTie Structure. The Bow-Tie structure is as shown in the figure shown here. As you can see, there are basically 3 sections: IN, SCC and OUT. No vertex from the SCC will direct a hyperlink to IN section. No vertex from OUT direct a user to SCC. SCC is the strongest connected core of the Web graph. In a sense, we can go from left to right i.e. IN-SCC, IN-OUT, SCC-OUT, but we can not trace back.



To find out SCC we used **networkx..strongly_connected_components(G).** Then for all the vertices in SCC, we find out their predecessors which are not in SCC. Those will be the set of vertices in IN section. Find out their successors which are not in SCC. Those will be the set of vertices in OUT section. Once we found the three sections, then we had an analysis as to what fraction of nodes in each of the three sections are essentially Netflix Originals or Prime Originals in the respective networks.

2.1. NETFLIX

Network	SCC	IN	OUT	In tendril	Out tendril	Tubes	Others
Netflix	918	3976	0	7	0	0	0
Netflix Originals	805	15	0	0	0	0	0
Fraction of Netflix Originals	0.877	0.0038	-		-	-	-

2.2. Amazon Prime

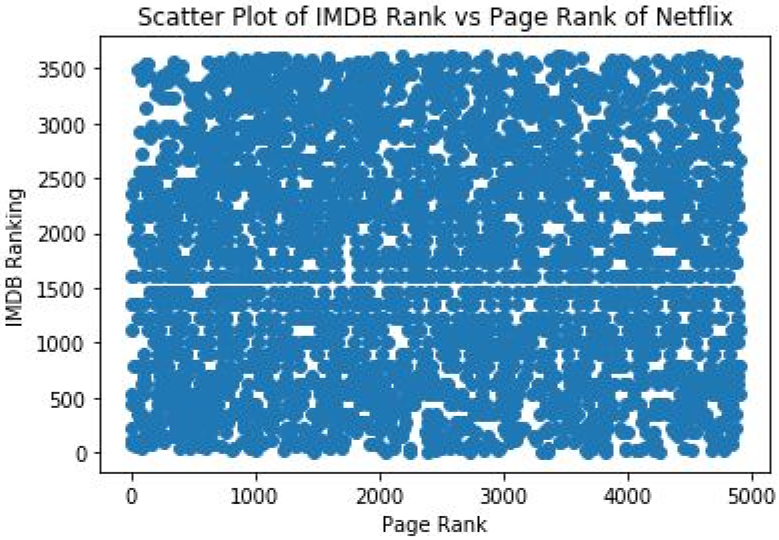
Network	SCC	IN	OUT	In tendril	Out tendril	Tubes	Others
Amazon	1	2092	0	176	0	0	0
Amazon Originals	0	41	0	1	-	-	-
Fraction of Amazon Originals	0	0.0196	0	0.0057	-	-	-

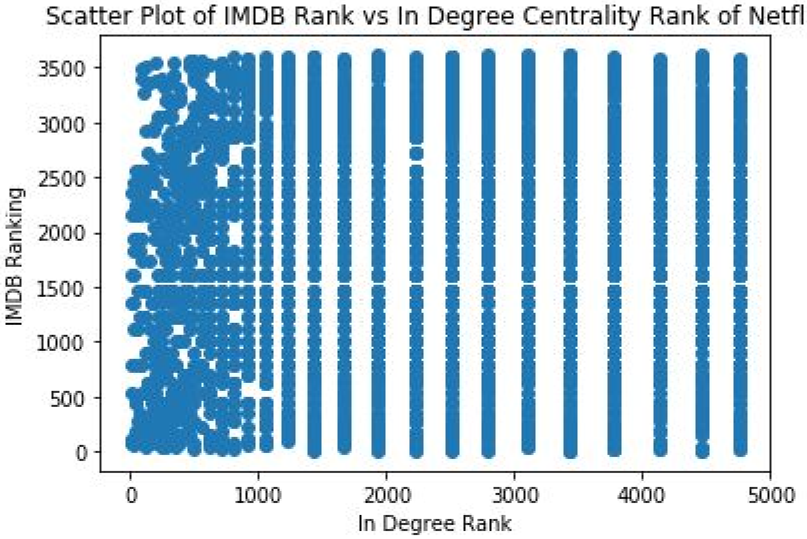
Following are some procedures done on the graph which will effectively help us to figure out whether there is any scandalous promotional bias or not.

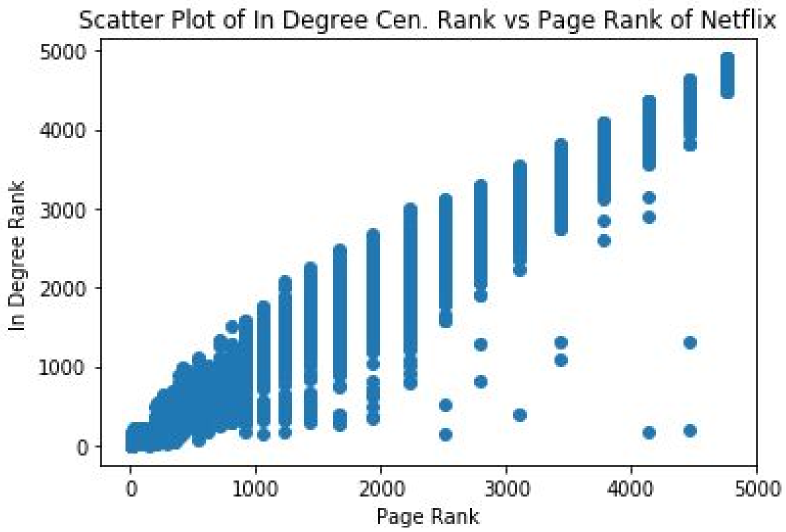
3. Scatter Plots of Rank vs. Rank

Here, first we found out the Pagerank and Indegree centrality ranks of all the nodes of the graph. Then these ranks were compared with IMDB ratings to check if Netflix/Amazon_Prime Recommendation System is biased. If it is found that the movies/shows with low IMDB ratings has higher Pagerank or Indegree centrality ranks derived from those Recommendation System, then this shows a symptom of biasness in those systems. Further we plot Pagerank Vs IMDB rankings and Indegree centrality rank Vs IMDB rankings for better understanding of the presence of biasness if any.

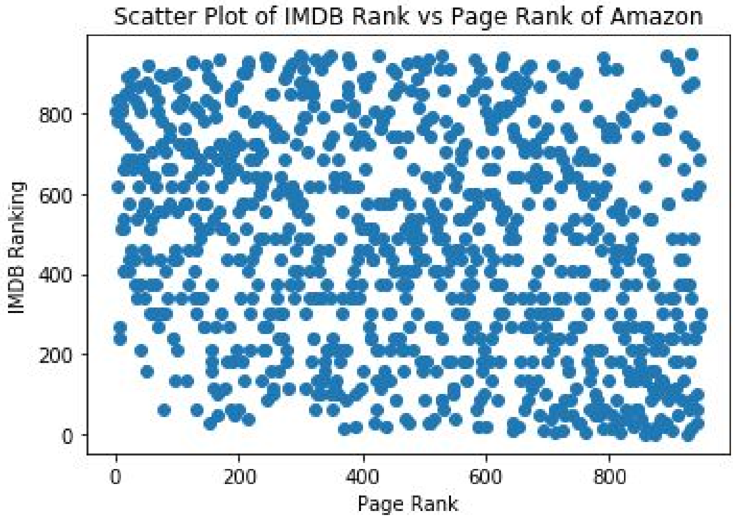
3.1. NETFLIX

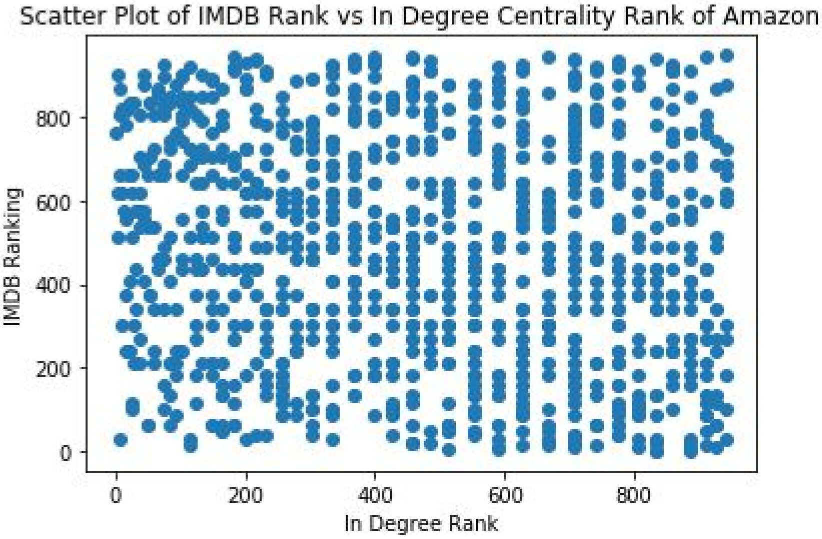




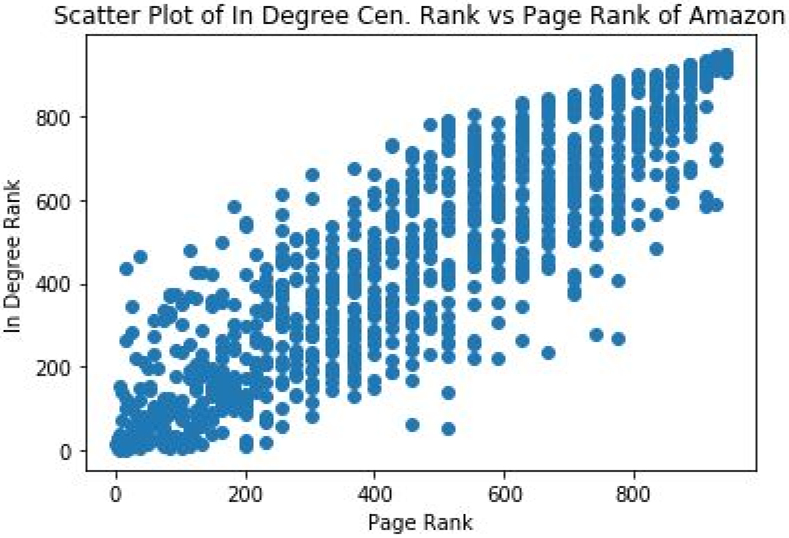


3.2. Amazon Prime





Scatter Plot of IMDB Rank vs In Degree Centrality Rank of Amazon



Scatter Plot of In Degree Cen. Rank vs Page Rank of Amazon

4. Spearman Rank Correlation Among Ranked Lists

Further from the above calculated ranks we evaluated the Spearman Rank Correlation between those Ranked lists. (the library used was **from scipy.stats import spearman**)

4.1. NETFLIX

Spearman Rank Correlation	**Netflix Pagerank**	**Netflix In Degree Cen Rank**	**IMDb Ratings**
Netflix Pagerank	1	0.973	0.027
Netflix In Degree Cen Rank	0.973	1	0.031
IMDb Ratings	0.027	0.031	1

4.2. Amazon Prime

Spearman Rank Correlation	**Amazon Prime Pagerank**	**Netflix In Degree Cen Rank**	**IMDb Ratings**
Amazon Prime Pagerank	1	0.876	-0.272
Amazon Prime In Degree Cen Rank	0.876	1	-0.138
IMDb Ratings	-0.272	-0.138	1

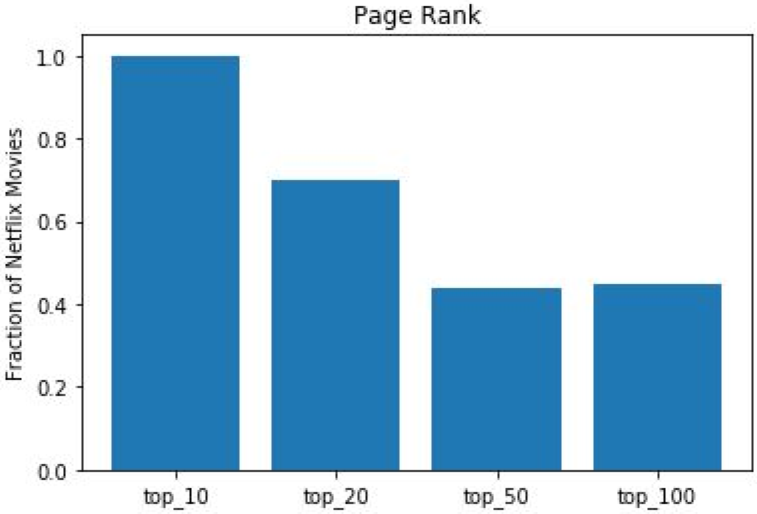
5. Distribution of Netflix Originals and Non-Netflix originals in Top-K

As we know, given we are ranking N number of movies as per their pagerank centrality And Indegree centrality measure, so essentially we are saying that these are the important nodes in this network. In our context, these are recommendations a particular movie is getting i.e. the promotion of the corresponding movie in the movie streaming platform. So, the distribution among these top ranked lists are important to have a good proportion of both the Netflix Originals and Non-Netflix Originals. To check the same, we took the Top-K ranked list and found out the number of Netflix originals and Non-netflix originals in each of the Top-K. Similarly for Amazon prime network. Then we took a note of the number of Netflix Originals or Prime Originals(x) respectively in Top-K in the following table. Basically K- x will be the number of Non-Netflix originals etc.

5.1. NETFLIX

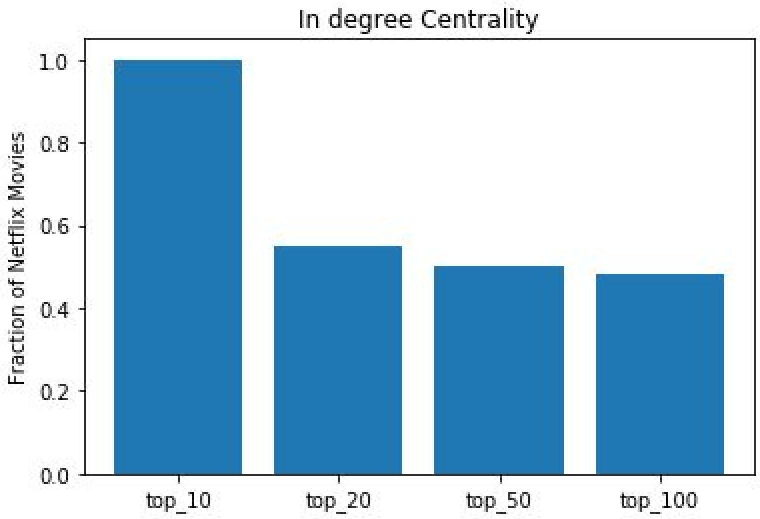
5.1.1. Pagerank

Network	**K=10**	**K=20**	**K=50**	**K=100**
Netflix	10	14	22	45



5.1.2. In degree Centrality Rank

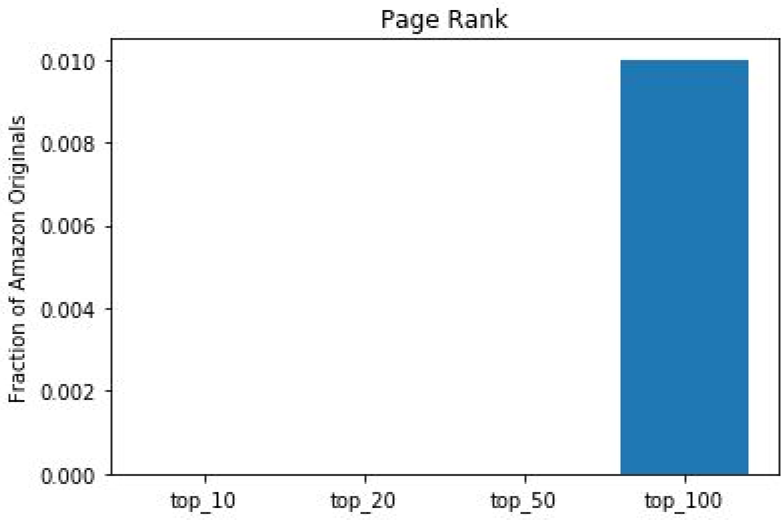
Network	K=10	K=20	K=50	K=100
Netflix	10	11	25	48



5.2. Amazon Prime

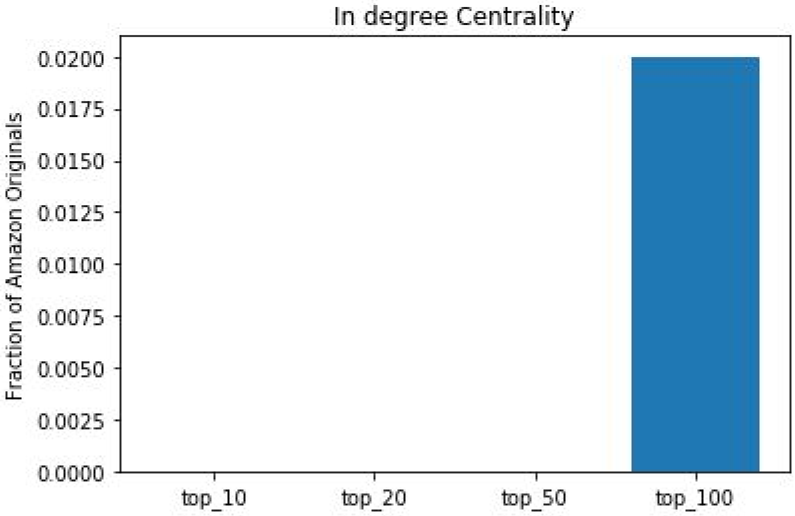
5.2.1. Page Rank

Network	K=10	K=20	K=50	K=100
Amazon Prime	0	0	0	1



5.2.2. In Degree centrality

Network	**K=10**	**K=20**	**K=50**	**K=100**
Amazon Prime	0	0	0	2



6. Bin Based Transition Contingency Table

We have two kinds of nodes. 1. Netflix Originals (Prime Originals) 2. Non-Netflix originals (Non-Prime Originals). For all the movies, we analyzed what is the fraction of recommendations going across different classes of nodes (movies). For e.g. in 6.1 the second number **4286** denotes the number of recommendations from movies of class Netflix Originals to movies of class Non-Netflix Originals.

6.1. NETFLIX

	Netflix Originals	**Non-Netflix Originals**
Netflix Originals	10405	4286
Non-Netflix Originals	21394	51551

6.2. Amazon Prime

	Prime Originals	**Non-Prime Originals**
Prime Originals	173	635
Non-Prime Originals	532	41979

Conclusion

The netflix data has been collected weekly to analyse how the recommendations are changing weekly but we have found that for such a small period of time all the properties of each week recommendations data were exactly same. Following points are the conclusion of each problems given in the project.

1. **Degree Distribution Log-Log Plot**: There are low number of nodes in both Netflix and Amazon Prime networks which have a very high value of Indegree. Most of the other nodes have a very low of indegree. One possible reason for this might be that the low number of nodes with high indegree are mostly Netflix originals or Amazon Prime Originals which could further explain the biasness of these recommendation systems by making all nodes in the networks to direct/recommend the user to the originals.
2. **Bow Tie Analysis:** None of the movie platform followed exact bow-tie-structure. However we can see that in netflix, the strongly connected components constitutes of 0.877 fraction of Netflix originals. The out section in bow tie structure of both the platform is equal to 0.
3. **Correlation Analysis:** We can conclude from the scatter plots and correlation values that there is no correlation between IMDb ranking and Page Rank or IMDb ranking and In degree centrality ranking for Netflix and Amazon prime. And hence we can say that there is biasness as they and recommending the movies that not according to popularity of that movie but there must be some other criteria.
4. **Top k- movies analysis:** We can clearly see that in case of Netflix that there is biasness as all the movies in top 10 are of their originals according to page rank and in degree centrality measure and approx 50% in top 20, 50 and 100 however prime videos are very less as it is a new platform so we can't comment about its biasness.
5. **Bin Based Transition Contingency Table:** In the case of Netflix network, the ratio of of total number of Netflix original nodes to the Non-Netflix nodes is quite low compared to the ratio of nodes pointing to Netflix orignals to that of node pointing to Non-Netflix nodes.

   This shows biasness in the Netflix network towards Netflix originals. In an unbiased case the ratio of of total number of Netflix original nodes to the Non-Netflix nodes should have been almost same as the ratio of nodes pointing to Netflix orignals to that of node pointing to Non-Netflix nodes.

   In case of Amazon Prime similar biasness is observed from the contingency table but with lower biasness.

Appendix

Codes and Datasets: click here