An Analysis of the Most Popular TV Shows

Maithreyi Venkatesh

11th January 2015

Abstract

In this project, an empirical analysis of the most popular TV shows on IMDB was carried out. Data from the Internet Movie Database (IMDB) was used to construct TV show networks whose key properties were analysed with respect to the data. The results showed that the shows that shared the most actors were either series' that had been running for close to 20 years or talk shows. Analysis of the top 10 highest weighted edges between two TV shows i.e. those that shared the highest number of actors also revealed the existence of clusters with the TV show network. It appears that talk shows share a high number of actors and cluster together while the "Law & Order" TV franchise cluster together.

1 Introduction

It was observed that many actors who had appeared in a TV show by one writer also appeared in other shows written by the same writer. This may also be the case for movies where an actor may star in multiple movies produced or directed by the same person. However because TV shows can run for multiple seasons where each season is made up of a number of episodes they have the potential to span multiple years. This could mean that two TV shows are likely, over the course of their seasons, to share far more actors than two movies.

Many researchers have carried out studies to analyse the movie and movie actor collaboration networks.[2][] Given the nature of TV shows, we can postulate that the TV show network may be very different from the movie network. However TV shows have never been studied in the context of complex networks. This project carried out an empirical analysis of the TV show network derived from data about the most popular TV shows.

To be more specific, this project aimed to investigate the research questions presented below. These research questions are described in more depth in section 3.

Research Question 1: What do the properties of the TV show network tell us about the most popular TV shows?
This research questions aims to empirically analyse the TV show network by extracting and relating key network properties to the TV show network in order to better understand it.

Research Question 2: To what degree do the characteristics of the TV show network change when all actors who have only appeared in only one episode in any show are omitted from the dataset?
As mentioned above, TV shows are commonly broken down into seasons of series' each of which comprises of a number of episodes. This means that the number of actors who appear in a given TV show is extremely large. Intuition tells us that the vast majority of actors may only appear in one episode. This research question aims to observe whether this is the case and if so, then to what extent does this affect the characteristics of the TV show network when the actors who only appear in a single episode are removed.

Research Question 3: Which shows share the most number of actors? Are there any specific reasons for this?

This research questions aims to determine which of the most popular TV shows share the most actors. In addition to this it also analyses of these connections between TV shows how many were trivial i.e. actors who only appeared in one episode and how many were significant? Finally, it investigates whether there were any specific reasons for a certain number of connections.

2 Data

The Internet Movie Database (IMDB), the most authoritative and comprehensive source on movies, television shows and actors was used in this project.

Python[1] and a scraping library, BeautifulSoup[2] , was used to acquire and process the data.

2.1 Data Acquisition and Dataset Description

Obtaining the necessary data for this task proved to be a difficult problem as IMDB does not have an official API. Unofficial APIs such as the Open Movie Database (OMDB) was investigated however it did not provide a full cast for each show. Additionally an IMDB ID was required to obtain the TV show details.

A list, entitled "Most Popular TV Series With At Least 5,000 Votes" [3], and the Internet Movie Database site was scraped for data about TV shows and their cast. It was assumed that these pages were kept up-to-date with a list of all the actors.

Only actors who were credited for their roles were retrieved from the pages. If actors, for some reason, were not credited then they did not need to be listed as part of the cast. This is similar to situations where if a researcher's name were not listed on a paper then it is unlikely that they contributed any significant work to it. Hence clearly either the actor did not wish to be credited or did not play a significant role in the TV show and was not used.

Two datasets were constructed from information gathered from the IMDB. The first, entitled shows_list, contained a list of the show. The second, entitled show_details, contained details of each of the TV shows from the previous list. A detailed description of both these datasets are provided below.

show_list contains:

1. imdb_id: The unique IMDB ID of the TV show.
2. name: The name of the TV show.
3. imdb_url: The URL to the official IMDB page of the TV show.
4. year: The year in which the TV show started and ended.

show_details contains:

1. imdb_id: The unique IMDB ID of the TV show.
2. name: The name of the TV show.
3. imdb_url: The URL to the official IMDB page of the TV show.
4. actor_id: The unique ID of the actor.
5. actor_name: The name of the actor who appeared in this TV show.
6. episodes: The number of episodes that the actor has appeared in the TV show. If this is unknown, yet the actor is credited, then this field contains "Unknown".

[1] `https://www.python.org/`
[2] `http://beautiful-soup-4.readthedocs.org/`

2.2 Dataset Analysis

In this section the results of simple analysis of the datasets acquired from the Internet Movie Database (IMDB).

The dataset contained 858 distinct TV shows. Though the IMDB page, described in 2.1, contained 859 at the time when it was scraped, one of the listings was a TV movie and was removed from the dataset.

The dataset also comprised was 140,378 actors in total. Given the small number of shows, this is an extremely large number of actors. This can be explained by the fact that TV shows run for many seasons and have multiple episodes per seasons. This means that an actor could be a part of the main cast, recurring cast (those who generally don't act in every episode) or simply appear in a single episode.

However there are many actors who play characters that only appear in a single episode. These actors can be deemed to be irrelevant to the overarching premise of the show. Removing any actors who had only appeared once in any of the shows left only the main and recurring cast. This dramatically reduced the number of actors and left only 51,375 actors who had appeared in more than one episode in at least one TV show.

Additionally, of the credited actors there were many for which the number of episodes in which they appeared for a given TV show was listed as "unknown". There were 1071 actors who had been credited for an "unknown" number of episodes.

3 Methodology

In this section an outline of methods used to answer each research question is provided. Gephi[3], the network visualisation and analysis tool was used to obtain the network measures. Python and NetworkX [4] was used to process and visualise the data when required.

Before providing a description of the methodology used to address the research questions, we first provide definitions of the key network properties used to characterise the TV show network. [4]

Average Node Degree: The degree of a node in a network is the number of links or edges has with another node. The average node degree or average degree of nodes denoted the average number of connections per node in the graph. This measure only takes into account the existence of a connection not the weight, if any, of it. [1]

Average Weighted Node Degree: This is the average weight of the connections per node in a weighted graph.

Modularity: The modularity is a measure of the strength of division of the network into communities or clusters. A high modularity indicates that the network has dense connections between the nodes within modules but sparse connections between nodes in different modules or communities.[6]

Average Clustering Coefficient: The clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. The average clustering coefficient is the average of the clustering coefficient of all nodes in the network.

Average Path Length: The average path length is the average distance between all node pairs in the network.[5].

[3] http://gephi.github.io/
[4] https://networkx.github.io/

3.1 Research Question 1

To answer the first research question, "What do the properties of the TV show network tell us about the most popular TV shows?", as outlined in section 1, a TV show network was constructed from the data.

TV Show Network 1 (TV-N1): This network was represented by a weighted, undirected graph. Each node in the graph represents a distinct TV show while a weighted edge between the nodes represents the number of actors that have acted in both shows. The highest the weight of the edge, the more actors the two TV shows shared.

The network made use of all 858 TV shows and 140,378 actors in the dataset to produce a graph with 858 nodes and 92474 edges. Following this, the properties described above were obtained from Gephi and analysed in the context of the TV shows.

3.2 Research Question 2

This research question aimed to assess the degree to which the TV show network properties were affected when all actors who have only appeared in a single episode from any TV show were omitted from the dataset.

Given the number of years that a TV show could span and the number of episodes per season, we can postulate that the majority of connections between TV shows due to actors who only appeared in a single episode. The data analysis, carried out in section 2.2, clearly showed that this assessment was valid. The dataset initially contained 140,378 different actors. Once actors who had not appeared in more than one episode in at least one TV show were omitted, the number of actors dropped down to 51,375 actors. This is less than half of the actors in the initial dataset indicating that most actors do only appear in a single episode.

To answer the research question and analyse the impact on the TV show of removing these actors, a second TV show network was constructed.

TV Show Network 2 (TV-N2): The second TV show network is also represented by an undirected graph with weighted edges. In this graph, each node represents a distinct TV show while an edge between nodes represents the actors who acted in both TV shows. The weight of an edge denotes the number of actors, who have acted in more than one episode, in both of the TV shows. Hence this network was constructed using 858 TV shows and 51,375 actors to produce a network with 858 nodes and 18043 edges.

Following the construction of these networks the key properties of these networks were obtained from Gephi for comparison in exactly the same manner as they were for the first research question.

3.3 Research Question 3

This research question aimed to determine which TV shows in the network shared the most actors. To answer this question the top 10 highest weighted edges in the TV-N1 and TV-N2 networks were extracted, visualised using the NetworkX Python library and analysed.

4 Results and Discussion

In this section we present the results of the experiments used to answer each of the research questions outlined in section 3 and discuss the results.

	TV Show Network 1 (TV-N1)
Average Node Degree	215.557
Average Weighted Node Degree	2718.21
Modularity	0.307
Average Clustering Coefficient	0.723
Average Path Length	1.83

Table 1: Key properties of the TV show network

4.1 Research Question 1

This research question aimed to characterise the TV show network by constructing a the TV-N1 network and extracting the key properties of this network, as described in section 3.1. The properties extracted are presented in Table 4.1.

The following statements can be made about the TV show network, TV-N1, using the values of the properties in Table 4.1:

- The average node degree indicates that on average each TV show is connected to 215 other TV shows. (It is impossible to be connected to 215.557 TV shows). This means that on average each TV show in the TV-N1 share an actor with 215 other TV shows.
- The average weighted node degree indicates that a TV show in TV-N1 network shares 2718 actors with another TV show in the network.
- The TV-N1 network has a low modularity, 0.307, indicating that the TV shows in TV-N1 do not share a large number of actors with other, specific TV shows.
- A high average clustering coefficient, 0.723, indicates that TV shows in TV-N1 shared actors with other TV shows and formed communities, or clusters, with other TV shows.
- An average path length of 1.83 means TV shows in the network are well connected and it is possible to move from one TV show another TV in the network in a low number of hops.

The low modularity but high average clustering coefficient appear to contradict each other. This could be attributed to the fact that some TV shows share incredibly large number of actors with each other. In analysing research question 3, whose results are presented in section 4.3, it was discovered that two TV shows in the network shared 1876 actors. In fact the top 10 highest weighted edges indicated that pairs of TV shows shared over 1000 actors in common. While the lowest 10 weighted edges in the network showed that pairs of TV shows only shared 2 or so actors.

4.2 Research Question 2

The second research question aimed to investigate the effect on the TV show network of omitting all actors who had only acted in one episode in any given TV show. From section 3.2 we recall that TV-N2 is an undirected, weighted graph just like TV-N1. However unlike TV-N1 it is made up of 858 nodes and 18,043 edges. The properties of the network are presented in Table 4.2. The properties of TV-N1 are also listed so that we can easily compare and contrast their results.

The following can be observed from the results:

- The average node degree of the TV-N2 indicates that on average each TV show in the new network is connected to 42 other TV shows as opposed to 215 other TV shows. This tells us that the majority of the TV shows did indeed share actors who only appeared in one episode in each.
- The average weighted node degree has decreased by more than a factor of 10 to 215. This indicates that a TV show in the TV-N2 network shares, on average, 215 actors with another TV show in the network.

	TV Show Network 1 (TV-N1)	TV Show Network 2 (TV-N2)
Average Node Degree	215.557	42.058
Average Weighted Node Degree	2718.21	215.713
Modularity	0.307	0.484
Average Clustering Coefficient	0.723	0.555
Average Path Length	1.83	2.393

Table 2: Statistical comparison of the TV show network with all cast members (TV-N1) versus the TV show network with main and recurring cast members only.

- A slight increase in the modularity of the TV-N2 network indicates that TV shows in this network share a large number of actors with other specific TV shows in the network when compared to TV-N1.
- A decrease in the average clustering coefficient that TV shows in TV-N2 do not form clusters to the same extend as the TV shows in TV-N1.
- An increase in the path of TV-N2 also indicates that the network is not as well connected as TV-N1. This is expected and was a direct result of the omission of all actors who had not acted in more than one episode in at least one TV show.

Additionally, it should be noted that the modularity and average clustering coefficient are much closer together in values. This means they are not contradicting each other as they did in TV-N1. This also confirms that this was caused by the incredibly number of actors who only acted in a single episode of any given TV show.

4.3 Research Question 3

By analysing the TV show network that included all TV shows and actors, this research question aimed to determine the pairs of TV shows that shared the most number of actors. It also aimed to investigated whether there were any specific reasons as to why two TV shows shared a certain number of actors.

The top 10 highest weighted edges in both TV-N1 and TV-N2 were extracted and visualised as subsections of the graphs.The results are presented in Figures 4.3 and 4.3, respectively. The mappings of node values are displayed in Tables 4.3 and 4.3, respectively. It should be noted that in both these tables the number of seasons directly corresponds to the number of years that the shows have been running for.

In TV-N1 "Law & Order: Special Victims Unit" (node 5) and "Law & Order" (node 7) shared 1876 actors. This was the highest number of actors shared by two TV shows in the entire dataset (including the actors who only acted in a single episode in any TV show). One explanation for this is that both shows are a part of the same franchise. Another explanation, as shown in Table 4.3, is that these shows have had 16 and 20 seasons, respectively. They are two of the longest running TV shows; "Law & Order: Special Victims Unit" started in 1999 and is still running and "Law & Order", the original, ran for 20 years from 1990 and ended in 2010.

The second highest number of actors of 1768 actors are shared by two talk shows; "Late Night with Conan O'Brien" (node 3) and "The Tonight Show with Jay Leno" (node 6). Both these shows been running for incredibly long periods of time, 17 years and 22 years to date, respectively. Additionally, the fact that both these are both talk shows is significant. Talk shows tend to have multiple guests per show and given that they are both talk shows, the same guests are likely to have appeared on both shows. These guests who appear on talk shows may not necessarily be actors they can be comedians, politicians, writers, music artists etc.

In fact, the TV shows listed in Table 4.3 are either a part of the Law & Order franchise which are long running TV shows or talk shows.

Node ID	TV Show Name and Year	No. of Seasons
1	The Late Late Show with Craig Ferguson (2005-)	11
2	Late Show with David Letterman (1993-)	22
3	Late Night with Conan O'Brien (1993-2009)	17
4	Jimmy Kimmel Live! (2003-)	13
5	Law & Order: Special Victims Unit (1999-)	16
6	The Tonight Show with Jay Leno (1992-)	22
7	Law & Order (1990-2010)	20
8	Law & Order: Criminal Intent (2001-2011)	10

Table 3: Mapping of node IDs, or labels, from graph in Figure 4.3 to names of TV shows

Node ID	TV Show Name and Year	No. of Seasons
1	The Late Late Show with Craig Ferguson (2005-)	11
2	Late Show with David Letterman (1993-)	22
3	Late Night with Conan O'Brien (1993-2009)	17
4	Jimmy Kimmel Live! (2003-)	13
5	Law & Order: Special Victims Unit (1999-)	16
6	The Tonight Show with Jay Leno (1992-)	22
7	Ellen: The Ellen DeGeneres Show (2003-)	12
8	Law & Order (1990-2010)	20

Table 4: Mapping of node IDs from graph in Figure 4.3 to names of TV shows

In TV-N2, the network where all actors who only acted in a single episode in all the TV shows were omitted altered the number of shared actors. None of the top 10 highest weighted edges in this network exceeded 800 actors with the highest being 799 actors shared between "Late Night with Conan O'Brien" (node 3) and "The Tonight Show with Jay Leno" (node 6). This is the same connection that appeared in TV-N1 as the second highest weighted edge.

As with Table 4.3, Table 4.3 also only contained the "Law & Order" franchise while the rest are all talk shows.

This indicates that the TV shows that share the most actors are either shows that have been running for a significant period of time or talk shows. Talk shows are not only long running but also have many guests per episode and over the course of year, these guests seem to appear more than once.

On a final note, we are able to identify two clusters in the graphs. In both graphs, we can see that the talk shows form one cluster (nodes 1, 2, 3, 4, 6 in Figure **??** and nodes 1, 2, 3, 4, 6 and 7 in Figure **??**) while the "Law & Order" franchise form another cluster (nodes 5, 7 and 8 in Figure **??** and nodes 5 and 8 in Figure **??**). This makes sense because the same guests are likely to have been invited to appear on multiple talk shows. It is surprising that none of the actors who appear in "Law & Order" have never appeared on any of the talk shows.

5 Further Work

The results of this study show that interesting information can arise from the analysis of the TV show network. The fact that TV show networks have never been analysed from a complex network perspective means that there is still a lot that can be done. Research into the TV actor collaboration network to analyse whether it is a small-world network or whether there is a a specific actor from which a similar "Bacon number" can be extracted are possible endeavours. Other options include a temporal analysis of the TV show network to analyse the evolution of the TV industry over the years.

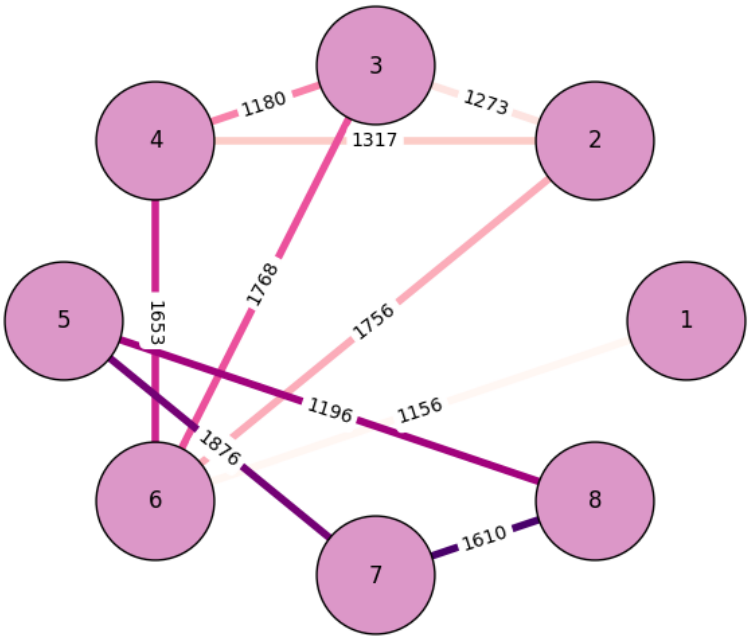

Figure 1: Subsection of the TV show network, TV-N1, made up of the top 10 highest weighted edges. Each node represents a TV show and each weighted edge represents the number of actors the TV shows share. The name, year and number of seasons each TV show in this graph can be found by mapping the node ID or label to the number in Table 4.3.

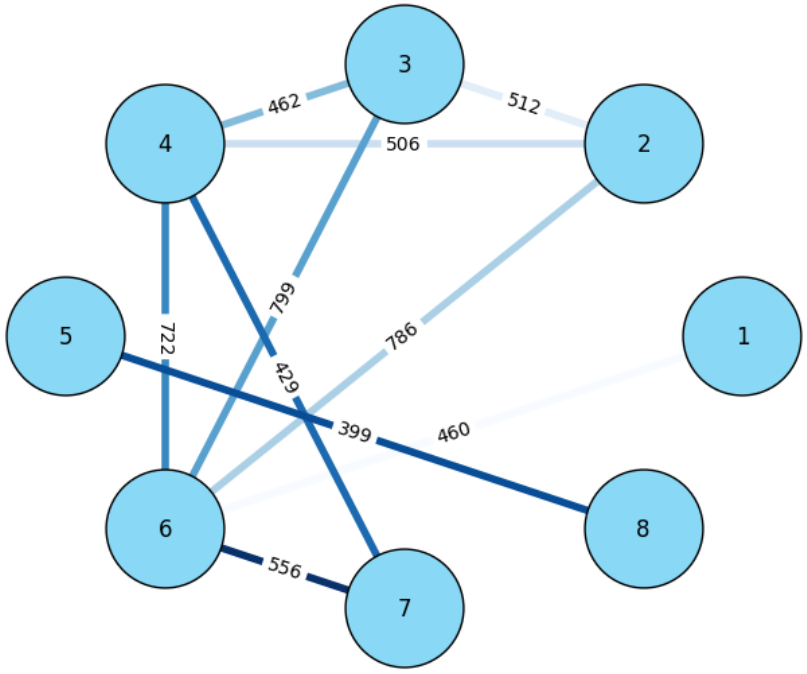

Figure 2: Subsection of the TV show network, TV-N2, made up of the top 10 highest weighted edges. Each node represents a TV show and each weighted edge represents the number of actors the TV shows share. The name, year and number of seasons each TV show in this graph can be found by mapping the node ID or label to the number in Table 4.3.

References

[1] Node degree. https://wiki.gephi.org/index.php/Degree. Accessed on 08-01-2015.

[2] A. Ahmed, V. Batagelj, X. Fu, S.-H. Hong, D. Merrick, and A. Mrvar. Visualisation and analysis of the internet movie database. Accessed on 01-01-2015.

[3] T. I. M. Database. Most popular tv series with at least 5,000 votes. http://www.imdb.com/search/title?num_votes=5000%2C&sort=moviemeter&title_type=tv_series. Accessed on 02-01-2015.

[4] J. Mahoney. Analysing network visualization statistics. http://coursedata.blogs.lincoln.ac.uk/tag/gephi/, jun 2012. Accessed on 08-01-2015.

[5] P. McSweeney. Average path length. https://wiki.gephi.org/index.php/Avg_Path_Length. Accessed on 08-01-2015.

[6] P. McSweeney. Graph density. https://wiki.gephi.org/index.php/Modularity. Accessed on 08-01-2015.