Community of the Top 50 Movies

Mehmet Ali Altunsoy
Computer Science Department
Faculty of Engineering
Bilkent University
Ankara, Turkey
ali.altunsoy@ug.bilkent.edu.tr

ABSTRACT
We all have movies that have been in our lives. In this research, we examine the relationships in the top 50 movies of IMDb [1] using network analysis. There are many studies on the connection between the words in the titles of the movies and the popularity of the movies [2][3][4]. What about the relationships between the actors who play in the movies and even the directors who shoot the movies? Using the data we get from IMDb, we will analyze the networks of the movies in the top 50, the networks of the actors in these movies, and the networks of the directors of the movies using community detection algorithms [5].

Keywords
Network of movies; Network of directors; Network of actors; Network analysis of movies; IMDb movie analysis.

1. INTRODUCTION
We watched movies many times on television, in cinema, and on digital platforms. We all have our favorite movies, directors, and actors. For most of us, following the relationships of these artists with each other in the magazine is also a different field of entertainment. Let's look at the relationships of artists in a more scientific way, leaving the magazine dimension aside, by examining the top 50 movies with the highest ratings of IMDb, the internet's most known and followed movie database.

One of the most important factors in making a movie is the director of the movie. The director decides what form the movie will take. Every director has a cast that he enjoys working with and that takes part in most of his movies. In fact, sometimes we see the same cast from the same director in movies with similar themes so much that the names of the movies are mixed up. Therefore, we can easily say that the thing that creates the movies is the director and actors of the movie. In this research, we will use this feature of the movies to examine the network of movies in the top 50 on IMDb by looking the common cast and directors, the network of directors who shot these movies based on the actors they work with, and the networks of the cast who acted together in these top 50 movies, using network analysis tools.

Applying community detection algorithms, clusters of the networks are detected. The features of these clusters and their relation will be discussed later in the paper. At the end of the research, we derive the connections of the directors and actors of these top rated 50 movies with create a meaningful network that contains their own communities.

2. BACKGROUND
Research has been done on the relationships between the words used in the titles of the movies and the popularity of the movies [6], examining the networks of the characters in the movies [7], and predicting the scores of the movies with neural network analysis [8]. The analysis of IMDb data is intriguing for a variety of reasons. For one thing, most people are familiar with and can relate to films and actors. The purpose of this research is to examine the communities established by successful movies themselves, their directors and actors, and see their closeness to each other, using network analysis.

3. DATASET
We collect our data from the IMDb. We download the data of the top 50 movies including their title, director, and cast with the IMDbPY package [9].

#	Attribute	Type	Description
1	Title	String	Title of the movie
2	Cast	String Array	List of cast act in the movie
3	Director	String	Director of the movie

Table 1: Attributes of Movies.

After getting the raw data. We manipulate them according to each network we are going to analyze. One of them is a network of movies, where movies are connected with respect to their director and their cast. Since, the directors' impact is much more than a single actor, the common directors' weight 4 times more than cast. Another one is the network of directors, where directors are connected according to the cast they work with, we have 36 different directors in the dataset. Finally, a network of the cast who starring more than once in the top 50 movies. They are connected according to their co-starring experience. In raw data, there were 3837 different actors/actresses. After filtering the ones with no connections' it reduces to 232.

4. METHODOLOGY
We made a network analysis of the top 50 movies of IMDb. We analyzed the communities in three different networks with two different community detection algorithms. There are several community detection algorithms in the network analysis area. In this research, we used two of them as mentioned. We choose these algorithms because modularity and betweenness metrics are significant for this project. We want to analyze the partitioning and bridging nodes. After clustering the networks, the result is represented with color based graph that different clusters have different colors. We expected to see different but similar results after applying these algorithms and we discussed why they are different.

For three networks, we applied two algorithms. Thus, we have six different graphs to analyze. We applied the following community detection algorithms.

4.1 Louvain Community Detection

Louvain is a two-phase unsupervised algorithm: modularity optimization and community aggregation. The second step begins once the first has been finished. Both will be carried through until the network has no more alterations and maximum modularity has been attained. Network modularity M is computed as follows:

$$M = \frac{1}{2m}\sum_{i,j}(A_{i,j} - \frac{k_i, k_j}{2m})\delta(c_i, c_j)$$

Where, m is the sum of all edge weights, $A_{i,j}$ is the edge weight between node i and j, k_i is the sum of the edge weights connected to node i, c_i is the community of node i, and δ is the Kronecker delta function [10] which $\delta(x, y) = 1\ if\ x = y, o.w. 0$. After all nodes are insert in a community, community aggregation is done [11]. We implemented a python script to output our data to CSV [12] format to import our data to Gephi [13] and apply the Louvain Algorithm. Movies with more common cast will be valuate more similar (If two movies have same director it will increase the similarity more than one common actor/actress. Hence, the weight of the common director is more than common actors/actresses as mentioned in Dataset section.)

4.2 Girvan-Newman Clustering

By gradually eliminating edges from the original network, the Girvan–Newman algorithm finds communities. The communities are the remaining network's connected components. The Girvan–Newman algorithm focuses on edges that are most probable "between" communities, rather than attempting to develop a measure that tells us which edges are most vital to communities [14]. The steps of the algorithm for detecting communities are as follows:

1. The network's betweenness of all existing edges is determined initially.
2. Remove the edge(s) with the maximum betweenness.
3. All edges affected by the removal have their betweenness recalculated.
4. Steps 2 and 3 are repeated until there are no more edges.

We used Gephi to apply the Girvan-Newman clustering algorithm with the data generated from implemented python code.

5. RESULTS

The network of movies, network of directors, and network of cast using Louvain community detection algorithm and Girvan-Newman clustering algorithm are generated using Gephi. The detailed metrics of the networks in Table 8.

5.1 Network of Movies

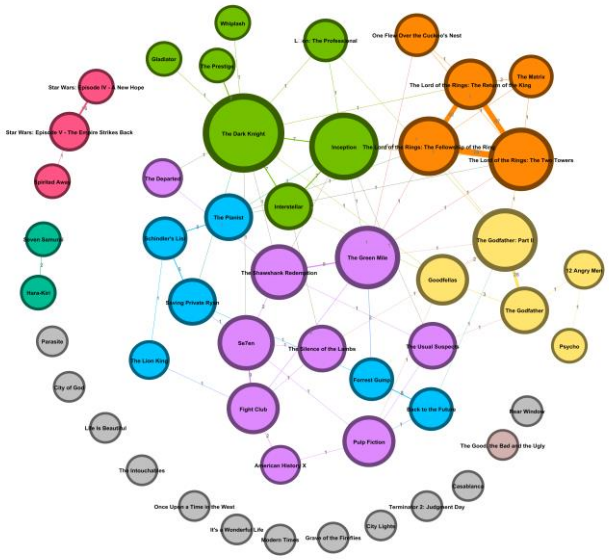

Figure 1: Network of Movies with Louvain Algorithm.

	%18	Crime Movies
	%14	Nolan Movies
	%12	Historical Movies
	%10	Mafia Movies
	%10	Lord of the Rings
	%6	Star Wars
	%4	Japanese Cinema
	%2	Indie Movies / Cult Classics

Table 2: Legend for Network of Movies with Louvain Algorithm.

The network of movies using Louvain clustering algorithm is generated as in Figure 1. The size of the nodes are based on their degree, and the colors of the nodes are according to their cluster detailed in Table 2. There are 20 communities in the network; however, 12 of them are movies with no connections. They are not connected to other movies mostly because of they are foreign, indie or old movies. The most crowded cluster is crime movies. This is followed by Christopher Nolan [15] movies, historical movies, mafia movies, Lord of the Rings series [16], Star Wars series [17], Japanese cinema, indie movies, and cult classics. From this network, we can number of actors and directors who work on similar genres are remarkable.

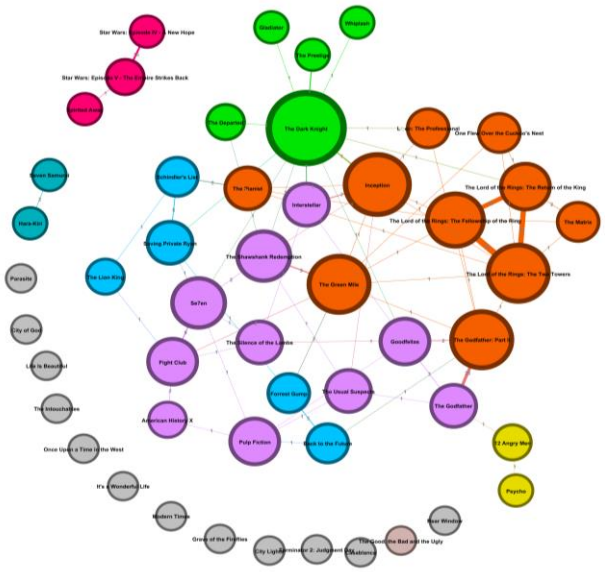

Figure 2: Network of Movies with Girvan-Newman Algorithm.

	%20	Crime Movies
	%20	Mixed
	%10	Historical Movies
	%10	Nolan Movies
	%6	Star Wars
	%4	Japanese Cinema
	%4	50’s Cinema
	%2	Indie Movies / Cult Classics

Table 3: Legend for Network of Movies with Girvan-Newman Algorithm.

The network of movies using Girvan-Newman clustering algorithm is generated as in Figure 2. The size of the nodes are based on their degree, and the colors of the nodes are according to their cluster detailed in Table 3. Statistics of the network is similar to the one with Louvain Clustering Algorithm. However, clustering is done differently. In the network there is a cluster contains movies from mixed genres from the Louvain clusters. This cluster contains some movies are the ones that become a bridge between other clusters, such as The Green Mile [18] and The Godfather Part II [19]. The reason behind the cast of these movies are the actors who can act in different genres.

5.2 Network of Directors

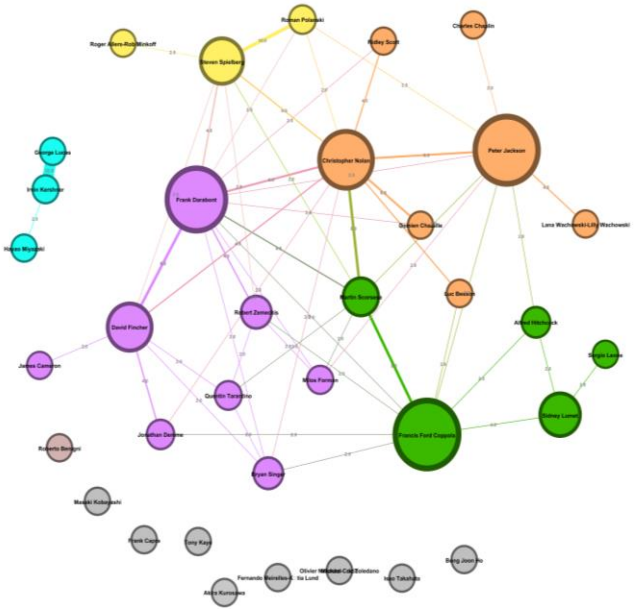

Figure 3: Network of Directors with Louvain Algorithm.

	%22	Mixed
	%19	Sci-Fi / Fantasy Directors
	%14	Mafia / Crime Directors
	%8	Spielberg - Polanski
	%8	Star Wars Directors
	%2	Foreign Directors

Table 4: Legend for Network of Directors with Louvain Algorithm.

The network of directors using Louvain clustering algorithm is generated as in Figure 3. The size of the nodes are based on their degree, and the colors of the nodes are according to their cluster detailed in Table 4. There are 14 communities in the network; however, 9 of them consist of single directors. The directors without any connection are mostly foreign directors. Hence, the actor network they can reach is limited compare to other directors in the graph. The rest of the graph consist of two components. One of these components consist of directors of Start Wars series, George Lucas [20] and Irvin Kershner [21]. Hayao Miyazaki [22] joins them with 2 common cast members. The cluster that have the most population is consist of mixture of directors. It is hard to classify them in one title. Concluded from the movies they took and the cast they work with, it can be said that, these directors are working on different genres and do not limit themselves. The second and third most crowded clusters consist of directors who mostly famous from one genre. For example, Peter Jackson [23] – Fantasy movies, Cristopher Nolan – Sci-Fi movies, Martin Scorsese [24] – crime movies, etc.

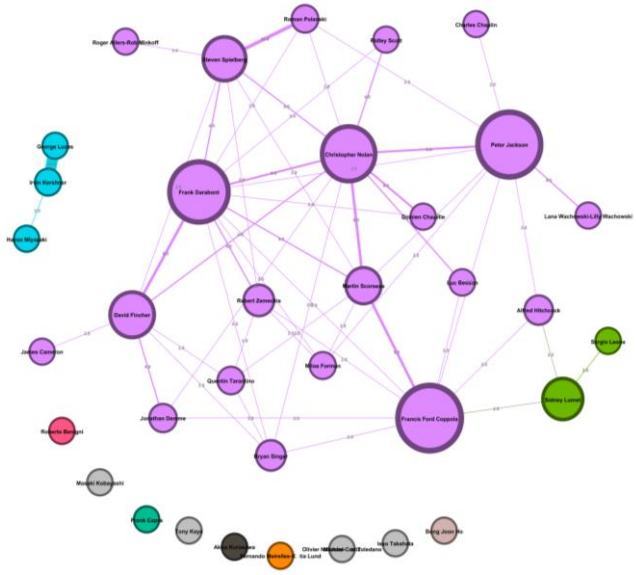

Figure 4: Network of Directors with Girvan-Newman Algorithm.

	%58	Mixed
	%8	Star Wars Directors
	%5	50's Directors
	%2	Foreign Directors

Table 5: Legend for Network of Directors with Girvan-Newman Algorithm.

The network of directors using Girvan-Newman clustering algorithm is generated as in Figure 4. The size of the nodes are based on their degree, and the colors of the nodes are according to their cluster detailed in Table 5. The number of clusters dramatically reduced with respect to Louvain clustering. This network express the community of directors in a more general way. Almost, all directors who is in the most populated cluster are directors from Hollywood. The rest of the network is, directors of Star Wars, 50's directors and foreign directors. It is interesting that even Star Wars movies are huge productions with many cast they have no common cast with the rest of the network.

5.3 Network of Cast

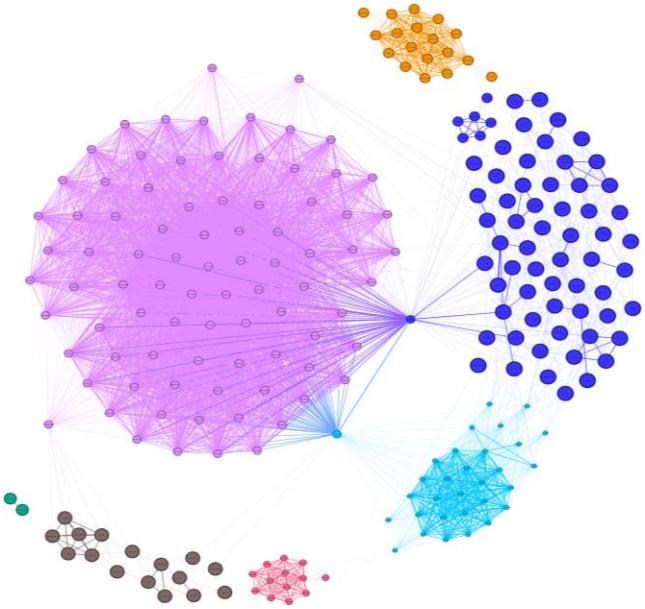

Figure 5: Network of Cast with Louvain Algorithm.

	%37	Cast of the Lord of the Rings Series
	%29	Mixed
	%13	Cast of the Godfather Series
	%8	Cast of the Star Wars Series
	%7	Chaplin and Hitchcock Cast
	%6	Cast of mid 20th century
	%1	Japanese

Table 6: Legend for Network of Cast with Louvain Algorithm.

The network of cast using Louvain clustering algorithm is generated as in Figure 5. The size of the nodes are based on their modularity class (just for distinguishing), and the colors of the nodes are according to their cluster detailed in Table 6. There are 7 communities in the network and 2 components. One of the component is consist of Japanese actors who act in the only two Japanese movies in top 50, Seven Samurai [25] and Spirited Away [26]. The rest of the network is mostly American actors.

The diameter of the graph is 8, thus an actor can reach to another actor with at most 6 people in between. It may more than expected however, the average path length is 2.7. Thus, we can say that it is a quite close network. The clusters in the network clearly separated from each other. The most crowded cluster is the cast of the Lord of the Rings series followed by mixed cluster (actors act in different genres), the cast of Godfather series, the cast of Star Wars series, cast of Charlie Chaplin and Alfred Hitchcock Cinema, and actors from mid-20th century.

There is three actors interesting in the network by looking, the ones who become bridge between two clusters. First one is Arnold Montey [27], who is an actors from Star Wars and Lord of the Rings

series. Even he play in side-roles he has the most connections in the network. Second one is Alan Lee [28], who is an Oscar winner conceptual designer and actor in Lord of the Rings series and act in a side role in the Godfather series. Finally, Luke Burnyeat [29] is an actor from Lord of the Ring series and played mini-roles in many famous movies such as, Dark Knight, V for Vendetta, Man of Steel, Skyfall, Shutter Island, etc.

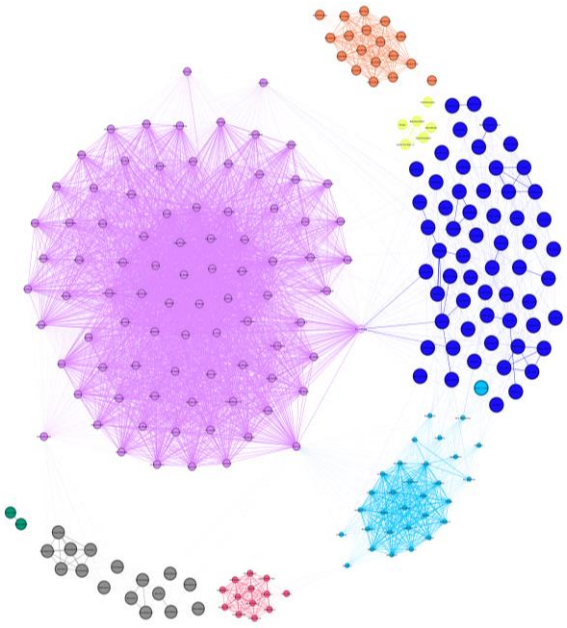

Figure 6: Network of Cast with Girvan-Newman Algorithm.

	%37	Cast of the Lord of the Rings Series
	%25	Mixed
	%13	Cast of the Godfather Series
	%8	Cast of the Star Wars Series
	%7	Chaplin and Hitchcock Cast
	%6	Cast of mid 20th century
	%3	Schindler's List & Pianist
	%2	Japanese

Table 7: Legend for Network of Cast with Girvan-Newman Algorithm.

The network of cast using Girvan-Newman clustering algorithm is generated as in Figure 6. The size of the nodes are based on their modularity class (just for distinguishing), and the colors of the nodes are according to their cluster detailed in Table 7. In the network, there is one more cluster than the one with Louvain clusters. The new cluster consist of the cast of Schindler's List [30] and Pianist [31]. An in-place addition, since the two movie have something in common and distinguishes from the rest of the network. Both movies about WWII and share many actors. The rest of the network is mostly same with the Louvain clustering.

Network	Movies		Directors		Cast	
Algorithm	L	GN	L	GN	L	GN
# of Communities	20	20	14	13	7	8
Modularity	0.597	0.38	0.597	0.1	0.341	0.374
# of Components	16		11		2	
Average Degree	2.880		3		41.597	
Nodes	50		36		232	
Edges	72		54		4867	
Diameter	6		5		8	
Average Path Length	2.496		2.20		2.768	
Density	0.059		0.086		0.182	

Table 8: Metrics of the Networks with Louvain (L) and Girvan-Newman (GN) algorithms.

6. CONCLUSION

We mention our conclusions about the networks we generated in specific in the previous section. On top of these, from all the networks we analyzed, the most significant inference is the remarkable amount of directors and actors have very characteristic styles. Their most successful works are based on similar genres. They have a comfort zone, they know what are they doing and do it in top level. Besides these, the other actors and directors are more flexible with what they are doing and it is hard to put them in specific classes. We clearly see and mentioned the examples of these actors and directors in our results.

7. REFERENCES

[1] "Top 250 Movies," Imdb.com. [Online]. Available: https://www.imdb.com/chart/top/. [Accessed: 06-Dec-2021].

[2] G. Bae and H.-J. Kim, "The impact of movie titles on box office success," J. Bus. Res., vol. 103, pp. 100–109, 2019.

[3] Augustine, Achal, and Manas Pathak. "User rating prediction for movies." Technical Report. University of Texas at Austin, 2008.

[4] Armstrong, Nick, and Kevin Yoon. Movie rating prediction. Technical Report, Carnegie Mellon University, 1995.

[5] T. D. Jayawickrama, "Community detection algorithms - towards data science," Towards Data Science, 29-Jan-2021. [Online]. Available: https://towardsdatascience.com/community-detection-algorithms-9bd8951e7dae. [Accessed: 06-Dec-2021].

[6] J.-M. Kim, X. Xiao, and I. K. Im, "Hollywood movie data analysis by social network analysis and text mining," Int. J. Electron. Commer. Stud., vol. 11, no. 1, pp. 75–92, 2020.

[7] P. F. Phalen, T. B. Ksiazek, and J. B. Garber, "Who you know in Hollywood: A network analysis of television writers," J. Broadcast. Electron. Media, vol. 60, no. 1, pp. 160–170, 2016.

[8] NASSER, Ibrahim M.; AL-SHAWWA, Mohammed O.; ABU-NASER, Samy S. A Proposed Artificial Neural Network for Predicting Movies Rates Category. 2019.

[9] "IMDbPY," Github.io. [Online]. Available: https://imdbpy.github.io/. [Accessed: 25-Oct2021].

[10] P. P. Camanho, A. Turon, and J. Costa, “Delamination propagation under cyclic loading,” in Delamination Behaviour of Composites, Elsevier, 2008, pp. 485–513.

[11] L. Rita, “Louvain Algorithm,” Towards Data Science, 09-Apr-2020. [Online]. Available: https://towardsdatascience.com/louvain-algorithm-93fde589f58c. [Accessed: 06-Dec-2021].

[12] K. Iqbal, “CSV File Format,” Fileformat.com, 10-Dec-2019. [Online]. Available: https://docs.fileformat.com/spreadsheet/csv/. [Accessed: 06-Dec-2021].

[13] “The open graph viz platform,” Gephi.org. [Online]. Available: https://gephi.org/. [Accessed: 06-Dec-2021].

[14] B. Kong, L. Zhou, and W. Liu, “Improved Modularity Based on Girvan-Newman Modularity,” in 2012 Second International Conference on Intelligent System Design and Engineering Application, 2012.

[15] R. B. H. Goh, Christopher Nolan: Filmmaker and philosopher. London, England: Bloomsbury Academic, 2021.

[16] P. Jackson, The lord of the rings: The fellowship of the ring. New Zealand, United States, 2001.

[17] I. Kershner, Star wars: Episode V - the empire strikes back. United States, United Kingdom, 1980.

[18] F. Darabont, The Green Mile. United States, 1999.

[19] F. F. Coppola, The Godfather: Part II. United States, 1974.

[20] G. Hansen, George Lucas: Cineasta y creador DE star wars / filmmaker and creator of star wars. Abdo Kids Jumbo, 2018.

[21] F. P. Miller, A. F. Vandome, and J. McBrewster, Eds., Irvin Kershner. Alphascript Publishing, 2010.

[22] J. Lenburg, Hayao Miyazaki. Chelsea House, 2012.

[23] B. Sibley, Peter Jackson: A film-maker’s journey. London, England: HarperCollins Entertainment, 2010.

[24] V. Lobrutto, Martin Scorsese: A Biography. Westport, CT: Praeger, 2007.

[25] A. Kurosawa, Seven Samurai. Japan, 1956.

[26] H. Miyazaki, Spirited Away. Turtleback Books, 2003.

[27] “Arnold Montey,” IMDb. [Online]. Available: https://www.imdb.com/name/nm7419291/?ref_=nv_sr_srsg_0. [Accessed: 30-Dec-2021].

[28] “Alan Lee,” IMDb. [Online]. Available: https://www.imdb.com/name/nm0496769/?ref_=nv_sr_srsg_3. [Accessed: 30-Dec-2021].

[29] “Luke Burnyeat,” IMDb. [Online]. Available: https://www.imdb.com/name/nm7572327/?ref_=nv_sr_srsg_0. [Accessed: 30-Dec-2021].

[30] S. Spielberg, Schindler’s List. United States, 1994.

[31] R. Polanski, The Pianist. United Kingdom, France, Poland, Germany, 2003.

[32] M. Altunsoy, Network-of-movies. [Online]. Available: https://github.com/malialtunsoy/network-of-movies. [Accessed: 30-Dec-2021]