Is this movie a milestone? Identification of the most influential movies in the history of cinema

Livio Bioglio and Ruggero G. Pensa

Abstract The success of a movie is usually measured through its box-office revenue or the opinion of professional critics, but such measures may be influenced by external factors, such as advertisement or trends, and are not able to capture the impact over time of a film. A more efficient measure should account to what extent a given movie has influenced other movies produced after its release, from both the artistic and the economic point of view. Hence, we propose a ranking method for movies based on the network of citations between them, obtained by combining several centrality indexes. We apply our method on a subset of the IMDb citation network consisting of around $65,000$ international movies, and we derive a list of films that can be considered milestones in the history of cinema. For each movie we also collect its year of release, genres and countries of production, and we analyze such features for finding trends and patterns in the film industry.

Key words: complex networks, network analysis, citation analysis, centrality, cinema

1 Introduction

Since its birth in the early 20th century, Cinema has played a huge role in culture, becoming recognized as the seventh art form [27]. Film industry has also a great importance in economy [2], becoming in recent years one of the most valuable economic resources for a country or a company. Thanks to his dual, and divergent, perception of cinema, as a product to sell or as a form of art, the success and importance of a movie is usually determined by either commercial or artistic criteria. Economical evaluation is based on the so-called box office, that is the amount of money raised by tickets sold in theaters and by home video market (sales and

Livio Bioglio e-mail: livio.bioglio@unito.it · Ruggero G. Pensa e-mail: ruggero.pensa@unito.it
University of Turin - Dept. of Computer Science, Turin, Italy

rentals), while artistic judgments are delivered by professional critics that evaluate the movie from an aesthetic and technical point of view. Both approaches have some limits. The financial performance of a film is related to other factors besides its quality, such as advertising, marketing expenditures, production cost, presence of star performers in the cast [24], to be a sequel or part of a bigger franchise [10], or trends driven by word of mouth [16, 20]. In addition, such approach makes it difficult to compare movies of different ages, even when box office revenue is adjusted for inflation[1], because these values may be inaccurate [1], and because sales may be influenced by economic performances and price level of the period when movie has been released [22]. Finally, the recent diffusion of video-on-demand technologies is changing the distribution of movies [31, 33], making the release on theaters less important for the success of a movie. From the artistic point of view, reviews are edited by human experts in the field of cinema, that may be subjective and may be biased by trends or ideologies. In addition, critics tends to focus on artistic merits, acclaiming movies with high aesthetic and intellectual level that result difficult to be appreciated by the majority of viewers, usually more interested in the entertainment aspect of a movie [18].

Here we propose a different metric of success for movies, an influence score related to how much a movie has influenced the ones produced after its release, that evaluates its importance in the history of cinema from both the artistic and the economic point of view. It is based on the inspiration that a film provides to others, even decades after its release, due to several factors, such as its style, creativity, innovation, story or franchise. Inspiration is evaluated through the network of citations obtained from a public available subset of IMDb, an online database of information about movies. Using this network we calculate the influence score for movies as a combination of four different centrality scores (in-degree, closeness, harmonic and pagerank centrality). We employ our score to rank the movies belonging to our dataset, with the aim of discovering the most influential movies in the history of cinema. In addition, we have also collected other features of each movie (year of release, genres and countries of production) and we analyze patterns and trends related to them as can be found in our dataset.

2 Related works

Networks of citations have been widely studied, especially in the domain of scientific publications, thanks to the huge amount of data available to researchers. In addition to analysis on their structures [28], such networks have been studied to find innovation trends [25, 5], and to quantify the impact of papers and authors [19, 23, 29], proposing measures widely used to estimate the scientific production, such as the Hirsch's h-index [17].

[1] an example of this calculation is performed by *Box Office Mojo*, available at http://www.boxofficemojo.com/alltime/adjusted.htm

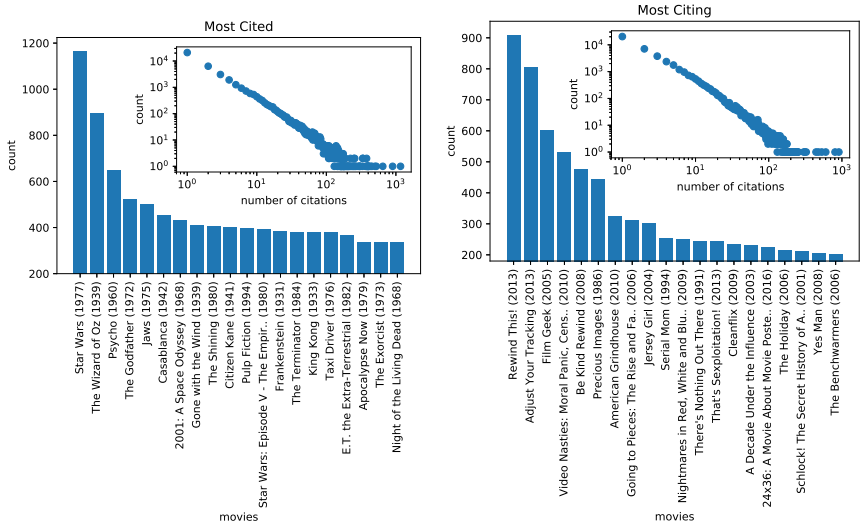
In the art domain, on the other side, it is rather difficult to infer a network of citations involving artists, because such relationships are not explicitly reported in an artistic product but must be inferred by human experts. Thus, researchers have focused their efforts on studying collaboration networks, in particular for music [14, 21] and cinema [3, 12, 13], where collaborations among musicians or actors are explicitly revealed. Nonetheless, some tentative has been done: in [11] the authors construct a creativity implication network of the visual art domain, using a computer vision algorithm to quantify similarity between artworks. For cinema, the availability of citation network produced by IMDb users has permitted to perform some research on this topic, in particular to infer the importance of a movie in the history of this art. In [32] the authors study the correlation between several metrics on movies (metacritic score[2], IMDb rating, box office, number of citations received, and pagerank score) and the presence of the movie in the US Library of Congress's National Film Registry (NFR), the United States National Film Preservation Board's (NFPB) selection of films for preservation in the Library of Congress, but their analysis are limited to movies produced in United States. In [8] the network of citations across movies furnished by IMDb is employed for studying the most inspiring movies and how their influence has evolved over years. The authors found that the inspiration of recent movies comes predominantly from the ones produced in the 70s and 80s, with some films from classical periods that still have a huge influence. However, their ranking of the most influential movies is rather straightforward, being simply the number of citations received. Finally, a closely related work is [30], where a combination of centrality metrics are employed for calculating an influence score for movies. This work, however, is limited by the fact that the centrality measure it proposes highly takes into consideration the temporal distance between citations, emphasizing the importance of most ancient movies. Furthermore, differently from our work, the analysis does not consider differences in genres, countries of production and years of release, but is only limited to computing the best ranked movies.

3 Citation network

IMDb is an online database of movies and TV series, featuring metadata such as year, country, genre, cast, production crew, budget, box office revenue, and so on. As of July 2017, IMDb includes information on more than 4 million titles (including episodes of TV series) and 8 million personalities (cast or crew)[3]. The site allows users to register for collaboratively expanding the database, by submitting new material, editing existing information and rating the movies stored in the database. IMDb also allows registered users to suggest connections between entries, indicating different kinds of relationships, from remakes to acknowledged source of inspi-

[2] Metacritic is a website that aggregates reviews of media products, showing an average score for each item. It is available at http://www.metacritic.com/.

[3] http://www.imdb.com/stats



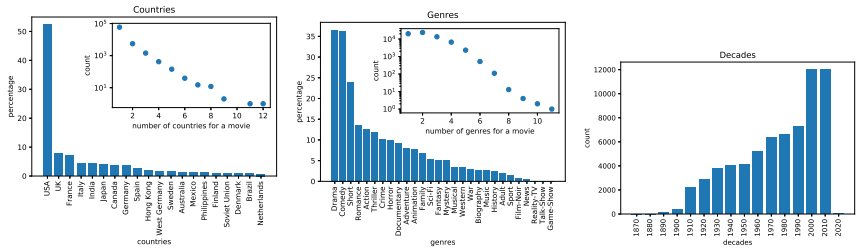
(a) Most cited movies (b) Most citing movies

Fig. 1 Most citing and cited movies in the citation network

ration. More in detail, there can be 6 different types of connection between a pair of movies (m_1, m_2): m_1 *features* m_2 if m_1 is mentioned or shown (for example on a screen or on a poster) in m_2; m_1 *references* m_2 if m_1 cites something appeared in m_2, such as a character, dialogue, fact, shot or iconic scene; m_1 *follows* m_2 if m_1 is a sequel of m_2, continuing its story; m_1 *spoofs* m_2 if m_1 contains some scene that is a parody of something involving m_2; m_1 is a *remake of* m_2 if m_1 is based on the same source of m_2; m_1 is a *spin off from* m_2 if m_1 tales a story that happens in the same narrative world of m_2, but it is not directly correlated to it. We employ these connections for constructing a network of citations between movies: in our idea, each connection can be seen as an influence made explicit by the director, the crew or the producers of the citing movie. Nodes of network are movies, and there exists a directed edge between two movies if there is at least one kind of connection, among the 6 recorded by IMDb dataset, between them.

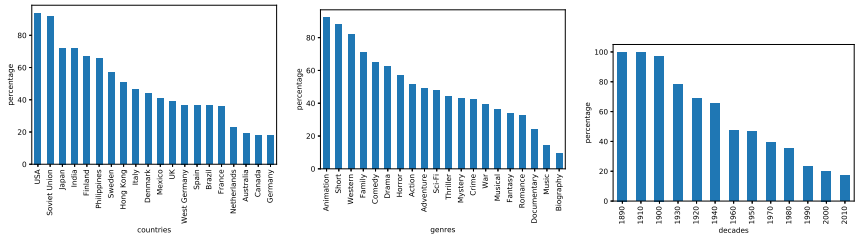
The network is extracted from the public available subset of IMDb[4]. In addition to citations, for each movie we also collect the year of release, the genres and countries of production. In our network we insert only movies having at least one citation, incoming or outgoing, a year of release, and at least one genre and one country specified. The final network is composed by $66,347$ nodes, representing movies, and $241,413$ edges, representing citations between movies; it is composed by $6,713$ weakly connected components (where the largest one is composed by the 71.3% of nodes and 79.5% of edges), and the average clustering coefficient of its undirected version is 0.194.

[4] available at http://www.imdb.com/interfaces



(a) Countries (b) Genres (c) Years of release by decade

Fig. 2 Number of movies by country of production, genre and year of release



(a) Countries (b) Genres (c) Years of release by decade

Fig. 3 Percentage of movies citing the same tag by country of production, genre and year of release

The distributions of degrees, both incoming and outgoing, and the titles of most cited and citing movies are depicted in Figure 1. The distributions of years of release (aggregated by decades), genres and countries are reported in Figure 2. We can notice that the majority of movies has been produced in the United States, and around 90% of the movies have been produced in only one country. The most represented genres are drama and comedy, roughly with the same amount of movies, and around 60% of films have been labeled with only one or two genres. Finally, the dataset mainly contains movies released after year 2000: it is not surprising, because from one side it shows the increase of film industry through ages, and on the other side IMDb is a crowdsourcing platform, in which users may be more interested in adding latest released movies respect to classical ones.

Finally, we analyze self-citations between groups. We select all the movies according to a given a tag, for example the genre "Drama", we look at the movies they cite, and then we examine how many times such movies are tagged as "Drama" themselves. Figure 3 depicts the percentage of self-citations in each tag of countries, genres and decades. We can notice an impressive percentage of self-citation for movies from United States, probably biased by the intrinsic unbalance of the dataset, and Soviet Union, maybe due to the cultural closure of this country, followed by Asian countries; on the other hand, Europe and South-America seem to be more open to influences coming from foreign countries. If we look at genres in Figure 3(b), we notice some important differences in self-citation attitude. Self-citations

are particularly significant in "Western" movies, that are focused on a certain kind of aesthetic widespread in the genre, "Short" movies, in which directors are usually more free to experiment techniques and styles, and "Animation" movies, technically more difficult to make and then produced only by few companies, that tends to reuse style, like Japanese Studio Ghibli, or characters, like Disney and Pixar studios. Finally, it is not surprising that movies tend to cite less other ones released in the same decade as time goes by, because more recent movies can combine techniques and ideas borrowed from more films respect the ones produced in the past.

4 Influence score

Here we describe how the influence score of a movie is calculated, by means of the citation network presented in section 3. It results from the combination of 4 different centrality scores: in-degree, closeness, harmonic and page rank. The in-degree centrality of a given node is simply the number of incoming edges. Closeness centrality [4] is calculated as the sum of the length of the shortest paths between a given node node and all other nodes in the graph. Let $d(x_1,x_2)$ be the distance function between two nodes x_1 and x_2, that calculates the length of shortest path from x_1 to x_2 (if x_2 can not be reached from x_1 the distance is 0): the closeness centrality for a generic node x is given by

$$C(x) = \frac{1}{\sum_{y \in N} d(y,x)}$$

Harmonic centrality [9, 26] for a node x is the sum of the reciprocal of shortest path distances from all other nodes to x, more formally:

$$H(x) = \sum_{y \in N} \frac{1}{d(y,x)}$$

Finally, Page rank centrality [7] is based on left dominant eigenvector, counting the number of possible ways any other node can reach x. This measure is well known because Google search engine is supposed to base the rating of web pages on it. For a given node x pagerank is defined as:

$$P(x) = d\boldsymbol{a}_x P(x) + \frac{(1-d)}{n}$$

where $P(x)$ is the pagerank centrality associated to node x, $d = [0,1]$ is the damping factor (the $1-d$ quantity is also known as restart probability), and $\boldsymbol{a}_x = [a_{1x}, \ldots, a_{nx}]$ is a vector such that each element $a_{ix} = 1/deg^+(x_i)$ ($deg^+(x_i)$ being the outdegree of x_i) if there exists a directed edge (x_i,x) ($a_{ix} = 0$ otherwise). For more details on centrality and centrality scores see [6].

Table 1 Top 20 movies by influence centrality

rank	title	countries	in degree	closeness	harmonic	pagerank	**influence**
1	The Wizard of Oz (1939)	US	0.774	1.0	1.0	1.0	**0.943**
2	Star Wars (1977)	US	1.0	0.772	0.8	0.706	**0.815**
3	Psycho (1960)	US	0.553	0.908	0.902	0.607	**0.742**
4	King Kong (1933)	US	0.32	0.922	0.885	0.544	**0.668**
5	2001: A Space Odyssey (1968)	US, UK	0.37	0.875	0.844	0.426	**0.629**
6	Citizen Kane (1941)	US	0.353	0.872	0.861	0.412	**0.625**
7	Metropolis (1927)	GER	0.126	0.966	0.933	0.417	**0.611**
8	The Birth of a Nation (1915)	US	0.0709	0.916	0.848	0.576	**0.603**
9	Frankenstein (1931)	US	0.329	0.808	0.802	0.411	**0.587**
10	Snow White and the Seven Dwarfs (1937)	US	0.179	0.877	0.835	0.398	**0.572**
11	Casablanca (1942)	US	0.388	0.739	0.761	0.368	**0.564**
12	Dracula (1931)	US	0.189	0.814	0.8	0.405	**0.552**
13	Nosferatu, eine Symphonie des Grauens (1922)	GER	0.156	0.844	0.801	0.36	**0.54**
14	The Godfather (1972)	US	0.452	0.67	0.676	0.351	**0.537**
15	Jaws (1975)	US	0.425	0.676	0.668	0.339	**0.527**
16	Cabiria (1914)	ITA	0.012	0.817	0.745	0.53	**0.526**
17	The Searchers (1956)	US	0.081	0.875	0.822	0.304	**0.521**
18	Bronenosets Potemkin (1925)	USSR	0.101	0.847	0.803	0.248	**0.5**
19	Dr. Strangelove or: How I Learned to...(1964)	US, UK	0.18	0.81	0.762	0.215	**0.492**
20	Gone with the Wind (1939)	US	0.352	0.633	0.654	0.31	**0.487**

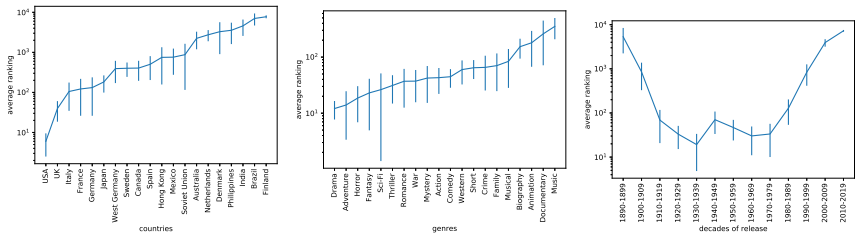
The ranking score for each centrality value is normalized, in order to obtain a value between 0 and 1, then the influence score for a movie is obtained by calculating the average of its four centrality scores. The entire workflow, from creation of citation network to calculation of influence score, has been performed by means of Python scripts[5], using the `networkx` library[6] [15] for calculating the centrality scores between nodes.

5 Analysis of the scores

Table 1 shows titles and centrality scores of the top 20 movies by influence score. As expected, several movies already appears in the list of most cited films in Figure 1(a), but in different positions: for example the most cited movie is in second position, while the ranking is lead by the second most cited movie, that has the best score in all the other centralities. Moreover, as expected from the distribution of countries in Figure 2(a), most movies in the list have been produced in the United States: only few movies in the list have been produced out from this country, and all of them belong to the early years of cinema. Finally, all the movies in the list have been released before 1980, and most of them even before 1940: such result is not completely surprising, because our score measures the influence of movies in history, then classical films, representing the first steps and experimentations in cinematic arts, have more probability of having influenced the following ones.

[5] available as jupyter notebooks at https://github.com/bioglio/most-influential-films

[6] available at https://networkx.github.io/.



(a) Countries (b) Genres (c) Year of release by decade

Fig. 4 Average ranking of top 10 movies by country of production, genre and year of release

Table 2 Top 10 movies by decade from 50's to 00's

	1950-59		1960-69		1970-79	
rank	title	global	title	global	title	global
1	The Searchers (1956)	17	Psycho (1960)	3	Star Wars (1977)	2
2	Vertigo (1958)	22	2001: A Space Odyssey (1968)	5	The Godfather (1972)	14
3	Singin' in the Rain (1952)	26	Dr. Strangelove or: How I Learned to... (1964)	19	Jaws (1975)	15
4	Touch of Evil (1958)	36	Dr. No (1962)	25	A Clockwork Orange (1971)	21
5	Shichinin no samurai (1954)	40	Night of the Living Dead (1968)	27	Taxi Driver (1976)	28
6	North by Northwest (1959)	43	Il buono, il brutto, il cattivo (1966)	29	The Exorcist (1973)	30
7	The Bridge on the River Kwai (1957)	60	Rosemary's Baby (1968)	44	The Texas Chain Saw Massacre (1974)	32
8	Rear Window (1954)	66	La dolce vita (1960)	54	Alien (1979)	57
9	Oklahoma! (1955)	78	Spartacus (1960)	56	Dirty Harry (1971)	69
10	Ben-Hur (1959)	80	Lolita (1962)	59	Apocalypse Now (1979)	70

	1980-89		1990-99		2000-2009	
rank	title	global	title	global	title	global
1	The Shining (1980)	38	Pulp Fiction (1994)	231	The Lord of the Rings: The Fellowship... (2001)	2353
2	Star Wars: Episode V... (1980)	41	Terminator 2: Judgment Day (1991)	365	Wo hu cang long (2000)	3200
3	The Terminator (1984)	55	The Silence of the Lambs (1991)	583	Gladiator (2000)	3389
4	E.T. the Extra-Terrestrial (1982)	93	The Matrix (1999)	681	Spider-Man (2002)	3671
5	Raiders of the Lost Ark (1981)	103	Reservoir Dogs (1992)	721	Kill Bill: Vol. 1 (2003)	4266
6	Scarface (1983)	129	Jurassic Park (1993)	742	X-Men (2000)	4292
7	A Nightmare on Elm Street (1984)	192	Goodfellas (1990)	1036	Charlie's Angels (2000)	4403
8	Cannibal Holocaust (1980)	194	Titanic (1997)	1167	Harry Potter and the Sorcerer's Stone (2001)	4519
9	The Evil Dead (1981)	199	Nikita (1990)	1309	Star Wars: Episode II... (2002)	4617
10	First Blood (1982)	243	True Romance (1993)	1578	Mission: Impossible II (2000)	4635

The list in Table 1, although interesting, does not offer an in-depth vision of the dynamical processes involving inspiration, because it compares movies of different genres, countries and years: for this reason we also analyze the ranking of movies in such coherent groups. To compare the influence of different classes we select the top 10 movies for each tag, then we calculate their average ranking in the complete list. The average ranks for each tag divided by country of production, genre and year of release are graphically summarized in Figure 4.

We start by looking at the most influential movies in each decade: the lists of the best movies from 1950 are reported in Table 2. From Figure 4(c) we can notice that movies released in the 30s are the most influential, followed by movies of the 20s, 60s and 70s, at approximately the same level. It is worth also discussing the list of top movies in range $2000-2009$: since they are too recent for having already influenced other movies, most of them belong to a big franchise (like the movies from "The Lord of the Rings", "Harry Potter" and "star Wars" series, or the superhero films), while the movie "Wo hu cang long", more known with its international title "Crouching Tiger, Hidden Dragon", is known for having introduced the so-called "wuxia" genre to Hollywood studios.

Table 3 Best 10 movies tagged as drama, comedy and horror

	Drama		Comedy		Horror	
rank	title	global	title	global	title	global
1	Citizen Kane (1941)	6	Dr. Strangelove or: How I Learned to...(1964)	19	Psycho (1960)	3
2	Metropolis (1927)	7	Singin' in the Rain (1952)	26	King Kong (1933)	4
3	The Birth of a Nation (1915)	8	La règle du jeu (1939)	37	Frankenstein (1931)	9
4	Frankenstein (1931)	9	The Wizard of Oz (1925)	46	Dracula (1931)	12
5	Casablanca (1942)	11	Sh! The Octopus (1937)	49	Nosferatu, eine Symphonie des Grauens (1922)	13
6	The Godfather (1972)	14	The Poor Little Rich Girl (1917)	50	Das Cabinet des Dr. Caligari (1920)	23
7	Jaws (1975)	15	The Tin Man (1935)	51	Night of the Living Dead (1968)	27
8	Cabiria (1914)	16	La dolce vita (1960)	54	The Exorcist (1973)	30
9	The Searchers (1956)	17	The Scarecrow (1920)	58	The Texas Chain Saw Massacre (1974)	32
10	Bronenosets Potemkin (1925)	18	Mr. Smith Goes to Washington (1939)	71	Un chien andalou (1929)	33

Table 4 Best 10 movies produced in France, Italy and Germany

	France		Italy		Germany	
rank	title	global	title	global	title	global
1	Un chien andalou (1929)	33	Cabiria (1914)	16	Metropolis (1927)	7
2	La grande illusion (1937)	35	Il buono, il brutto, il cattivo (1966)	29	Nosferatu, eine Symphonie des Grauens (1922)	13
3	La règle du jeu (1939)	37	La dolce vita (1960)	54	Das Cabinet des Dr. Caligari (1920)	23
4	La dolce vita (1960)	54	L'année dernière à Marienbad (1961)	74	Triumph des Willens (1935)	45
5	L'année dernière à Marienbad (1961)	74	8½ (1963)	86	Der Golem, wie er in die Welt kam (1920)	102
6	8½ (1963)	86	Per un pugno di dollari (1964)	110	Der Sieg des Glaubens (1933)	162
7	Étoile sans lumière (1946)	213	Per qualche dollaro in più (1965)	145	Tag der Freiheit - Unsere Wehrmacht (1935)	206
8	À bout de souffle (1960)	224	C'era una volta il West (1968)	152	Der Golem (1915)	245
9	Pépé le Moko (1937)	244	Cannibal Holocaust (1980)	194	Das Testament des Dr. Mabuse (1933)	255
10	Un témoin dans la ville (1959)	259	Un témoin dans la ville (1959)	259	Der blaue Engel (1930)	256

Figure 4(b) shows the top 25 average ranking of genres. We can notice that dramatic movies are more influential than comedies, even if the number of movies belonging to these two genres is approximately the same, as shown in Figure 2(b): such behavior is even more evident if we look at the list of top dramatic movies by genre in Table 3, in which all films in drama list belong to the global top 20 ranking, compared to only one film in the list of comedies. Another surprising result is the huge influence of horror movies, in particular if we refer to the list of horror films in Table 3: some of them have been released in the 20s or 30s, but several other ones have been released in the 60s and 70s, and show a global rank lower than 35. A possible explanation of this behavior is that horror movies tends to be less original and more influenced by trends, replying stories and situations of success in this genre.

Finally, we analyze the average ranking of top 10 movies released in a country, depicted in Figure 4(a). As expected, Anglo-Saxon movies lead the ranking, with the United States and the United Kingdom in the first two positions. Three countries are closely ranked around the third position: Italy, France and Germany. But if we analyze the year of release of movies in the top 10 list of these countries, reported in Table 4, we can notice three different behaviors: French movies span from the 30s to the 60s, while Italian ones are almost entirely clustered in the 60's, except for "Cabiria", which introduced several innovations to the cinematic language, like the moving camera, and "Cannibal Holocaust", an horror movie that creates the so-called "found-footage" genre. On the other hand, German movies are gathered in the 20s and 30s, highlighting the importance of this country during the early years of cinema. Finally, Indian movies seems to be not so influential in the history of cinema, despite the fact that, in our dataset, the number of movies produced in this country is similar to the number of movies produced in Italy, as shown in Figure 2(a).

Table 5 Presence in NFR and AFI lists of top 25 movies produced in United States

rank	title	NFR	AFI	rank	title	NFR	AFI
1	The Wizard of Oz (1939)	Yes	Yes	14	The Searchers (1956)	Yes	Yes
2	Star Wars (1977)	Yes	Yes	15	Dr. Strangelove or: How I Learned to... (1964)	Yes	Yes
3	Psycho (1960)	Yes	Yes	16	Gone with the Wind (1939)	Yes	Yes
4	King Kong (1933)	Yes	Yes	17	A Clockwork Orange (1971)	**No**	Yes
5	2001: A Space Odyssey (1968)	Yes	Yes	18	Vertigo (1958)	Yes	Yes
6	Citizen Kane (1941)	Yes	Yes	19	Tarzan the Ape Man (1932)	**No**	**No**
7	The Birth of a Nation (1915)	Yes	Yes	20	Singin' in the Rain (1952)	Yes	Yes
8	Frankenstein (1931)	Yes	Yes	21	Night of the Living Dead (1968)	Yes	**No**
9	Snow White and the Seven Dwarfs (1937)	Yes	Yes	22	Taxi Driver (1976)	Yes	Yes
10	Dracula (1931)	Yes	**No**	23	Il buono, il brutto, il cattivo (1966)	**No**	**No**
11	Casablanca (1942)	Yes	Yes	24	The Exorcist (1973)	Yes	**No**
12	The Godfather (1972)	Yes	Yes	25	Chang: A Drama of the Wilderness (1927)	**No**	**No**
13	Jaws (1975)	Yes	Yes				

5.1 Validation

Validation of our rank is a critical issue. Spitz and Horvát in [30] observe that their ranking score is not related to the number of awards a movie received, and the same happens when considering box-office revenue. We may compare our ranking with the ones provided by professional critics, but the latter are biased by the knowledge and opinion of each individual. Aggregating several human professional critics, like the Rotten Tomatoes website[7], can not be a solution, because our ranking states the influence of a movie, while critics rank movies according to other features, such as artistic or technical merits. As a partial solution, we compare our ranking only for movies produced in United States, with two lists of significant movies established by experts. The first one is the US Library of Congress's National Film Registry[8] (NFR), that contains 703 movies produced in United States that are "culturally, historically, or aesthetically significant". The second one is the list of the 100 best American movies, as determined by the American Film Institute (AFI) ranking: indeed, our list consists of 122 movies, resulting from the union of the ranking compiled in 1998 and 2007 [9]. Only 3 movies are ranked similarly in 1998 and 2007: such difference suggests that ranks made by human experts are highly biased, depending on trends and personal opinions of the authors and of the time when they are compiled. However, we can not simply compare these lists with our list of top movies, because they are not designed with the same purpose. NFR also contains historically significant movies, such as the first experiments of the use of color and sounds in movies, or videos that document habits or historical events, like the great San Francisco earthquake of 1906 or the footage that caught the assassination of the American President John F. Kennedy. On the other hand, AFI list contains the best movies, while here we analyze influence, that are surely related concepts but not exactly identical. For this reason, we decide to restrict our research only on the top 25 movies produced in US: Table 5 shows the list of movies and their presence in the two lists. We observe that among them, 21 are inserted in the NFR list and 19 in

[7] available at https://www.rottentomatoes.com

[8] available at https://www.loc.gov/programs/national-film-preservation-board/film-registry/complete-national-film-registry-listing/

[9] the rank of 1998 is availabe at http://www.afi.com/100years/movies.aspx, while the rank of 2007 is available at http://www.afi.com/100years/movies10.aspx

the AFI one. Among the movies excluded by the lists, some of them are principally produced in other countries, while some other ones are horror movies, a particular genre that is not liked by all the critics: a special exception is *Chang: A Drama of the Wilderness (1927)*, that is present in our top 25 list because it is referenced in the highly influential series of movies on the character of King Kong.

6 Conclusions

In this paper we have analyzed the network of citations between movies available from the official repository of the IMDb movie dataset. Combining four centrality scores (in-degree, closeness, harmonic and pagerank) we have calculated an influence score for each movie in the dataset, obtaining for each movie a ranking based on the citations received from other ones. In addition, we have combined other information on single movies (year of release, genres and countries of production) for analyzing how movies with different features have influenced the history of cinema. We have found that the most influential movies are the ones released in the 30s, almost equally followed by the ones released in the 20s, 60s and 70s. Drama, adventure and horror movies seem to be more influential than other genres, in particular when compared to comedy movies, despite the fact that the latter is one of the most widespread genres. Finally, Anglo-Saxon movies are the most influential in the history of cinema, as expected from the dataset, consisting of 50% of movies produced in the United States, followed by France, Italy, that has been influential in the 60s, and Germany, whose film production has been crucial in the early years of cinema.

References

1. Anderson, S.E., Albertson, S., Shavlik, D.: How the motion picture industry miscalculates box office receipts. Inland Empire Business Journal **25**(11), 16–24 (2003)
2. Bakker, G.: The economic history of the international film industry. Eh. Net Encyclopedia of Economic History (2005)
3. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. science **286**(5439), 509–512 (1999)
4. Bavelas, A.: Communication patterns in task-oriented groups. The Journal of the Acoustical Society of America **22**(6), 725–730 (1950)
5. Bioglio, L., Rho, V., Pensa, R.G.: Measuring the inspiration rate of topics in bibliographic networks. In: Proceedings of DS 2017 (2017). To appear
6. Boldi, P., Vigna, S.: Axioms for centrality. Internet Mathematics **10**(3-4), 222–262 (2014)
7. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems **30**(1), 107–117 (1998)
8. Canet, F., Valero, M.Á., Codina, L.: Quantitative approaches for evaluating the influence of films using the imdb database. Comunicación y Sociedad **29**(2), 151 (2016)
9. Dekker, A.: Conceptual distance in social network analysis. Journal of Social Structure (JOSS) **6** (2005)

10. Dhar, T., Sun, G., Weinberg, C.B.: The long-term box office performance of sequel movies. Marketing Letters **23**(1), 13–29 (2012)
11. Elgammal, A., Saleh, B.: Quantifying Creativity in Art Networks. In: Proceedings of the Sixth International Conference on Computational Creativity (ICCC 2015), pp. 39–46. Brigham Young University (2015)
12. Eom, Y.H., Jeon, C., Jeong, H., Kahng, B.: Evolution of weighted scale-free networks in empirical data. Physical Review E **77**(5), 056,105 (2008)
13. Gallos, L.K., Potiguar, F.Q., Andrade Jr., J.S., Makse, H.A.: Imdb network revisited: unveiling fractal and modular properties from a typical small-world network. PloS one **8**(6), e66,443 (2013)
14. Gleiser, P.M., Danon, L.: Community structure in jazz. Advances in complex systems **6**(04), 565–573 (2003)
15. Hagberg, A.A., Schult, D.A., Swart, P.J.: Exploring network structure, dynamics, and function using networkx. In: Proceedings of the 7th Python in Science Conference, pp. 11–15 (2008)
16. Hennig-Thurau, T., Wiertz, C., Feldhaus, F.: Does twitter matter? the impact of microblogging word of mouth on consumers' adoption of new movies. Journal of the Academy of Marketing Science **43**(3), 375–394 (2015)
17. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences of the United States of America **102**(46), 16,569 (2005)
18. Holbrook, M.B.: The Role of Ordinary Evaluations in the Market for Popular Culture: Do Consumers Have "Good Taste"? Marketing Letters **16**(2), 75–86 (2005)
19. Kaur, J., Ferrara, E., Menczer, F., Flammini, A., Radicchi, F.: Quality versus quantity in scientific impact. Journal of Informetrics **9**(4), 800–808 (2015)
20. Liu, Y.: Word of mouth for movies: Its dynamics and impact on box office revenue. Journal of Marketing **70**(3), 74–89 (2006)
21. Park, J., Celma, O., Koppenberger, M., Cano, P., Buldú, J.M.: The social network of contemporary popular musicians. International Journal of Bifurcation and Chaos **17**(07), 2281–2288 (2007)
22. Pautz, M.C.: The decline in average weekly cinema attendance, 1930-2000. Issues in political economy **11** (2002)
23. Petersen, A.M., Fortunato, S., Pan, R.K., Kaski, K., Penner, O., Rungi, A., Riccaboni, M., Stanley, H.E., Pammolli, F.: Reputation and impact in academic careers. Proceedings of the National Academy of Sciences **111**(43), 15,316–15,321 (2014)
24. Prag, J., Casavant, J.: An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry. Journal of Cultural Economics **18**(3), 217–235 (1994)
25. Renoust, B., Claver, V., Baffier, J.F.: Flows of knowledge in citation networks. In: International Workshop on Complex Networks and their Applications, pp. 159–170. Springer (2016)
26. Rochat, Y.: Closeness centrality extended to unconnected graphs: The harmonic centrality index. In: Applications of Social Network Analysis (ASNA 2009) (2009)
27. Sadoul, G.: Histoire du cinéma mondial : Des origines à nos jours. Flammarion (1976)
28. Sinatra, R., Deville, P., Szell, M., Wang, D., Barabási, A.L.: A century of physics. Nature Physics **11**(10), 791 (2015)
29. Sinatra, R., Wang, D., Deville, P., Song, C., Barabási, A.L.: Quantifying the evolution of individual scientific impact. Science **354**(6312), aaf5239 (2016)
30. Spitz, A., Horvát, E.Á.: Measuring long-term impact based on network centrality: Unraveling cinematic citations. PloS one **9**(10), e108,857 (2014)
31. Wasko, J.: The death of hollywood: Exaggeration or reality? The handbook of political economy of communications pp. 305–330 (2011)
32. Wasserman, M., Zeng, X.H.T., Amaral, L.A.N.: Cross-evaluation of metrics to estimate the significance of creative works. Proceedings of the National Academy of Sciences **112**(5), 1281–1286 (2015)
33. Zhu, K.: Internet-based distribution of digital videos: the economic impacts of digitization on the motion picture industry. Electronic Markets **11**(4), 273–280 (2001)