Correlations Between User Voting Data, Budget, and Box Office for Films in the Internet Movie Database

Max Wasserman
Department of Engineering Sciences and Applied Mathematics, Northwestern University, 2145 Sheridan Road, Room M426, Evanston, IL 60208. E-mail: maxwasserman2014@u.northwestern.edu

Satyam Mukherjee
Department of Chemical and Biological Engineering, Northwestern Institute on Complex Systems (NICO), Northwestern University, Chambers Hall, 600 Foster Street, Evanston, IL 60208. E-mail: s-mukherjee@kellogg.northwestern.edu

Konner Scott and Xiao Han T. Zeng
Department of Chemical and Biological Engineering, Northwestern University, Chambers Hall, 600 Foster Street, Evanston, IL 60208. E-mail: xiaohanzeng2014@u.northwestern.edu

Filippo Radicchi
Center for Complex Networks and Systems Research, School of Informatics and Computing, Indiana University, 919 10th Street, Bloomington, IN 47408. E-mail: filiradi@indiana.edu

Luís A. N. Amaral
Department of Chemical and Biological Engineering, Department of Physics and Astronomy, Northwestern Institute on Complex Systems (NICO), Howard Hughes Medical Institute (HHMI), Northwestern University, 2145 Sheridan Road, Room E136, Evanston, IL 60208. E-mail: amaral@northwestern.edu

The Internet Movie Database (IMDb) is one of the most-visited websites in the world and the premier source for information on films. Similar to Wikipedia, much of IMDb's information is user contributed. IMDb also allows users to voice their opinion on the quality of films through voting. We investigate whether there is a connection between user voting data and economic film characteristics. We perform distribution and correlation analysis on a set of films chosen to mitigate effects of bias due to the language and country of origin of films. Production budget, box office gross, and total number of user votes for films are consistent with double-log normal distributions for certain time periods. Both total gross and user votes are consistent with a double-log normal distribution from the late 1980s onward while for budget it extends from 1935 to 1979. In addition, we find a strong correlation between number of user votes and the economic statistics, particularly budget. Remarkably, we find no evidence for a correlation between number of votes and average user rating. Our results suggest that total user votes is an indicator of a film's prominence or notability, which can be quantified by its promotional costs.

Introduction

In today's world, we are seemingly in constant connection to the Internet. Most of our activities are stored in electronic databases, and the aggregation of this information represents a novel source for the study of human behavior (Castellano, Fortunato, & Loreto, 2009). Indeed, researchers have reported on the statistical properties of the communication patterns of e-mail (Ebel, Mielsch, & Bornholdt, 2002; Malmgren, Stouffer, Motter, & Amaral, 2008; Radicchi, 2009) and traditional "snail" mail (Malmgren, Stouffer, Campanharo, & Amaral, 2009), on the analysis of the macroscopic features of web surfing (Gonçalves & Ramasco, 2008; Johansen, 2001), and so on. Aggregate electronic information has not only been a boon to scientific investigation but also has demonstrated utility in

Received October 30, 2013; revised December 11, 2013; accepted December 11, 2013

 • Published online 21 May 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23213

many practical applications. For instance, aggregate information is used by eBay (ebay.com) to quantify the reputation of sellers and buyers, and researchers have used data from Twitter (twitter.com) to analyze collective moods (Golder & Macy, 2011) and monitor the spread of ideas (Aral, Muchnik, & Sundararajan, 2009).

The information present on the web is the result of the aggregation of the work of many individuals as well as the outcome of complex and self-organized interactions between large numbers of agents. For example, the vast amount of information contained on Wikipedia (wikipedia.org) is the product of millions of user contributions. The collaborative and collective outcome is not merely the sum of the knowledge of each individual contributor but also the result of continuous modifications and refinements by users. The content generated through this collaborative strategy is generally more complete than those produced by individuals because the collaborative framework ensures more control of the quality of the provided information (Woolley, Chabris, Pentland, Hashmi, & Malone, 2010; Wuchty, Jones, & Uzzi, 2007).

The Internet Movie Database (IMDb; http://www.imdb.com), one of the most frequently accessed websites worldwide (Alexa, 2013c), is home to the largest digital collection of metadata on films, television programs, videos, and video games. Similar to Wikipedia, IMDb's content is updated exclusively by unpaid, registered users. In addition to accepting user-contributed information, IMDb also allows users to rate, on a scale of 1 (worst) to 10 (best), the quality of any film or program. Adding new information to the database is a mostly altruistic activity, as it requires action on the part of the contributor to enhance the understanding of others. However, voting on the quality of a film is a less altruistic action because the user is able to voice his or her opinion through voting. Is there any useful information to be drawn from online voting information? IMDb is not a new subject for scientific analysis. It has been used in the context of studying actor-collaboration networks (Amaral, Scala, Barthelemy, & Stanley, 2000; Herr, Ke, Hardy, & Börner, 2007; Watts & Strogatz, 1998) and in developing recommendation systems (Grujić, 2008). More recently, IMDb's extensive collection of keywords was used to create a metric of film novelty (Sreenivasan, 2013). However, little work has been done on the connections between the user contributions to IMDb's database of information and user contributions to a film's rating score.

In this article, we identify correlations between IMDb's user ratings of films and various other characteristics reported in the database. To accomplish this, we use information available on IMDb to construct a directed network of films. This network provides the data set for our analysis. We proceed to filter the network to account for biases on the part of users. Using metadata collected on our filtered data set, we determine the distributions of various characteristics in several time windows to identify temporal changes. In addition, we perform linear regressions with the goal of quantifying correlations between user voting data and user-contributed information.

Data

We limit our analysis of IMDb data to films, including in our study both feature-length and short films. We chose to include short films because excluding them would ignore almost all films made before 1920. We retrieved data from IMDb on October 26, 2012, and therefore only include in our analysis films released by 2011. Note that we do not exclude documentary films.

The metadata on films included in IMDb consist of year of release, country of production, primary language, user voting statistics, and several types of financial information (Figure 1). All metadata are user edited, apart from the voting data, for which IMDb automatically tallies the total number of user votes and reports an average rating using an "in-house formula." From among all types of financial information reported for a film, we focus on production budget, box office gross in the United States, and greatest amount grossed in a single week during a film's theatrical run. All of these values are in unadjusted U.S. dollars and thus not corrected for inflation or gross domestic product (GDP) growth.

Among the plethora of available and editable information is a section titled "connections," a list of references and links between films and other media. All connections listed on IMDb are classified as one of eight "types": references, spoofs, features, follows, spin-offs, remakes, versions, and edits. We only consider the connections that are classified as references, spoofs, or features. A *reference* is a connection between films where one contains an homage to the other in some form. For example, the famous flying bicycle scene in *E.T.: The Extra-Terrestrial* (1982) is a reference to a sequence in *The Thief of Bagdad* (1924) where characters also fly in front of the moon. References also come in the form of similar quotes, similar settings, or similar filming techniques. A *spoof* is a connection between films where one mocks the other. For example, the wagon circle scene from *Blazing Saddles* (1974) is a spoof of the final scene from *Stagecoach* (1939). A *feature* occurs when a film includes an extract from another film. The scene in *When Harry Met Sally* (1989) . . . where the title characters watch *Casablanca* (1942) is an example of a feature connection. Our analysis is limited to these connections because they pertain to parts of films, such as scenes or quotes, and thus are conscious choices on the part of creative people such as directors, actors, and screenwriters.

Using the connections between films, we construct a network where each film is a node and where each connection is an arc (Figure 2). An arc connecting Movie A to Movie B indicates that Movie A contains a reference to Movie B (or spoofs Movie B or features a clip from Movie B). We admit a connection into the network only if the citing film was released in a later calendar year than was the cited film; that is, all links in our network are directed backward in time, and the network contains no links between two films

DOI: 10.1002/asi

FIG. 1. Typical metadata on films included in IMDb. Examples of metadata collected for *Raiders of the Lost Ark* (top panel). A depiction of two directed edges in the connections network. One edge represents a citation by *Raiders of the Lost Ark* to *Yojimbo* (The opening shot of the former honors the opening shot of the latter.) The other edge represents a citation linking *Home Alone* to *Raiders* (Villains in both films suffer burns to the hand that leave an impression.) (bottom panel). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

released in the same calendar year. This constraint ensures that the network is acyclic. The network we construct from the metadata on film connections consists of 32,636 films and 77,193 connections.

We limit our analysis to the largest weakly connected component of the network formed by film connections. This *giant component* consists of 28,743 films (88% of all films in the network) linked by 74,164 arcs (96% of all connections). For each film, we take note of its number of incoming and outgoing arcs. In network theory, these values are known, respectively, as the *in-degree* and *out-degree*. It is desirable that "connections" is a relatively recondite category in IMDb, as the presence of reported connections functions as a minimum threshold for consideration in our analysis. In addition, we assume that any film with a nonempty connections section has sufficient information in sections that are better known, such as box office information.

Biases in Metadata Reporting

Because IMDb is user edited, we must investigate possible biases in reporting. Although user editing allows a reference website such as IMDb to be up-to-date, it diffuses the responsibility for fact-checking, leading to greater uncertainty about accuracy and objectivity of information. Biases may be due to the makeup of the user base. For example, Wikipedia recently reported that 91% of its user editors are male (Wikimedia Foundation, 2011), which explains why female-focused topics are less thoroughly covered. Therefore, we evaluate the basic properties of a database to account for biases prior to performing a comprehensive analysis.

Two characteristics of a film that could reveal biases of the user editors are country of production and primary language. Thus, we assign films in the giant component to one of three groups based on country and language. Films

DOI: 10.1002/asi

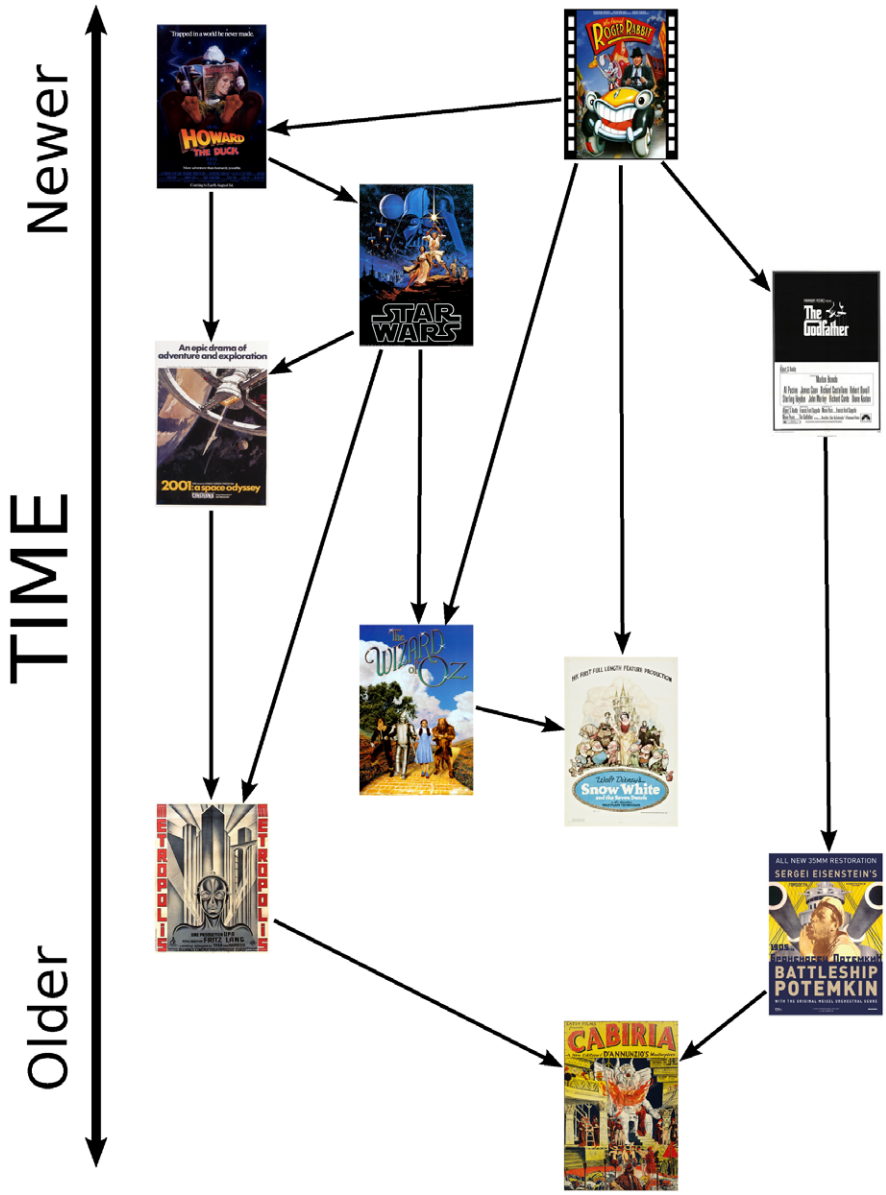

FIG. 2. Subgraph of the film-connections network. A subgraph containing 10 films of the 28,743 in the giant component of the film-connections network. Films are ordered chronologically, based on year of release, from bottom to top. A connection between two films means that a sequence, a sentence, a character, or another part of the referenced film has been adopted, utilized, or imitated in the referencing film. For example, there is a connection from *Star Wars* (third from the top) to *Metropolis* (third from the bottom) because C-3PO is modeled on the robot from Fritz Lang's 1927 film. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

produced in the United States, regardless of language, are assigned to the "USA" group. Films made outside the United States with English as the primary language are assigned to the "English non-USA" group. Films produced outside the United States and in a language other than English are designated as the "non-English non-USA" group. Note that the USA group comprises a majority of the films in the connections network.

In Figure 3, we consider the time dependence of various properties for each group. The number of new films released annually increases over time for each of the three groups, particularly rising during the last two decades (Figure 3A). The decrease in new films from 2009 to 2011 was presumably caused by the recession of 2008 to 2010, but it also may be due to a reporting delay for low-budget films.

There is a stark difference in the number of votes films receive depending on their year of release and grouping (Figure 3B). USA films released after 1990 have a median of more than 2,000 user votes, whereas those released before 1980 have a median below 500. In addition, there is a trend-reversing dip in the median number of votes received by films beginning around 1995. We attribute this reversal to a sizable jump between 2003 and 2008 in the number of new

DOI: 10.1002/asi

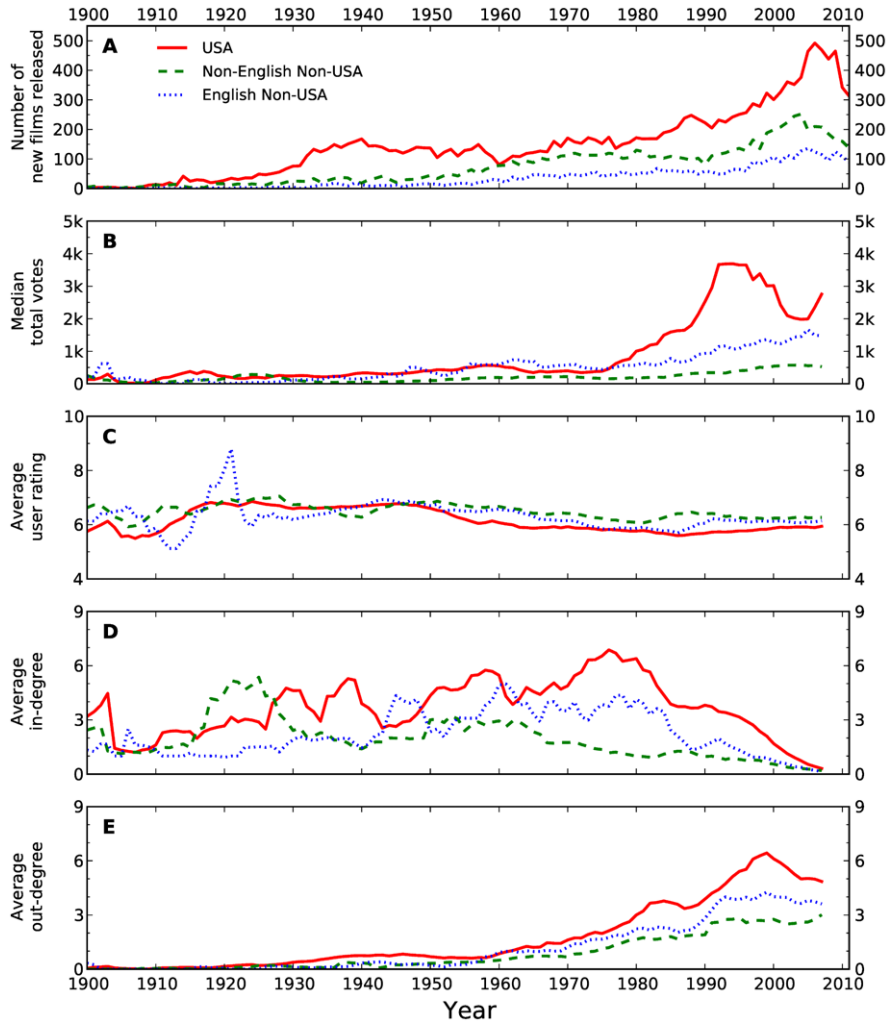

FIG 3. Characteristics of movies in the giant component. We partitioned films in the giant component of the connections network into three groups: USA films, English non-USA films, and non-English non-USA films. (A) Time dependence of number of films released annually. Time dependence over 5-year windows of (B) median number of votes, (C) average user ratings, (D) average in-degree, and (E) average out-degree. We use 5-year windows to calculate all statistics apart from number of films released because of data variability on a year-to-year basis. We show the median for (B) because the relatively few films with large numbers of votes (i.e., ≥100,000) heavily skew the mean, making it less representative of a typical film. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

films released, which we presume were mostly independently produced, "obscure" pictures that likely receive few votes on IMDb.

Films in the English non-USA group receive more votes than do films from the non-English non-USA group, despite non-English non-USA films outnumbering English non-USA films by almost two to one (Figure 3B). Similarly, we observe that the English non-USA group averages more incoming and outgoing citations than does the non-English non-USA group, despite the latter being more numerous in the data set (Figure 3D,E). These findings suggest that there is both a language bias and a temporal bias in the distribution of user votes in IMDb. These biases are not observed in the average user ratings for films, as there is little change over time and among the three groups (Figure 3C).

Surprisingly, the average in-degree declines for films released after 1992 while the average out-degree increases for films released after the mid-1980s (Figure 3D,E). The latter is to be expected because connections between films can only travel "backwards" in time (i.e., from a newer film to an older film). Unexpectedly, the average out-degree declines for films made after 1999, particularly for films in the USA group.

The presence of a temporal bias toward recent films in the IMDb connections network is not unexpected, as modern technology allows films to be produced more quickly and potentially less expensively than ever before. This trend in costs makes the downward trend in average out-degree for films beginning around 1999 stand out. It is unlikely that films released in 1998 are more citation laden than are films

DOI: 10.1002/asi

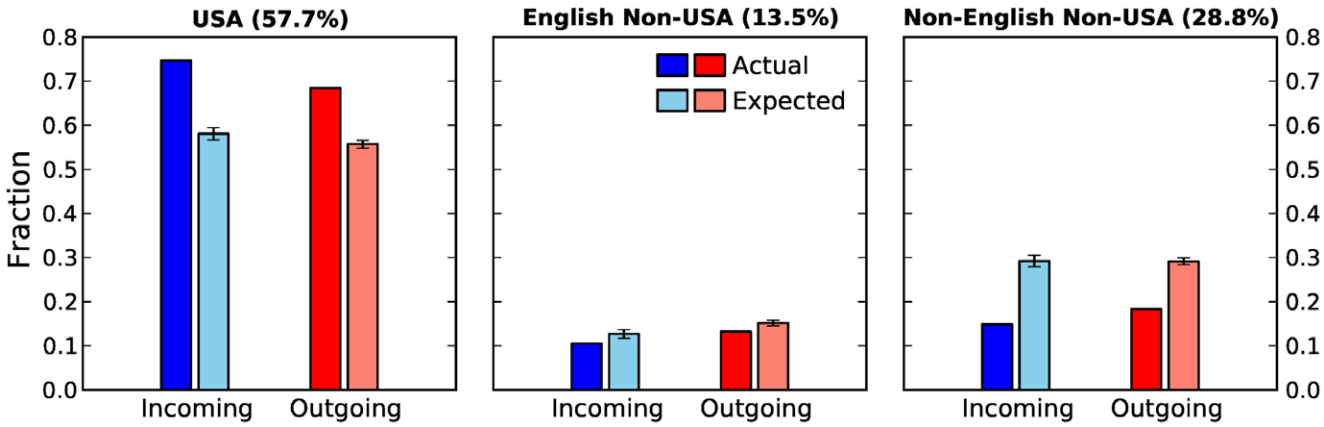

FIG. 4. Actual and expected fractions of connections by country/language grouping. Fractions of both incoming and outgoing connections in the giant component of the film-connections network for USA films (left plot), English non-USA films (center), and non-English non-USA films (right). Numbers in parentheses represent the percentages of nodes in the network that belong to each group. Dark blue and dark red bars represent the fractions of connections in the giant component. Light blue and light red bars represent the average fraction of connections in the giant component following a Monte Carlo simulation where the films were randomly reassigned to one of the three country/language groupings. Error bars represent 1 *SD* of the mean following 10,000 simulations. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 1. Bias due to country of production and primary language. Note how USA films mostly reference other USA films (i.e., 84.5% of connections point to 57.7% of the nodes) while mostly ignoring non-English non-USA films (i.e., 6.7% of the connections point to 28.8% of the nodes).

		%Links from		
Group	% of nodes	USA	English non-USA	Non-English non-USA
USA	57.7	84.5	64.8	45.1
English non-USA	13.5	8.8	21.8	8.4
Non-English non-USA	28.8	6.7	13.4	46.5
Total	100.0	100.0	100.0	100.0

released in 2005. Instead, we presume that this decrease indicates a different sort of temporal bias wherein IMDb does not yet have full information for films made in the 2000s, whereas information for films made in the 1990s is more complete. Thus, the data suggest that there is a latency period for the reporting of film connections, a fact that must be taken into consideration in interpreting any analysis.

To better quantify biases in enumeration of connections, we consider the observed proportions of incoming and outgoing connections for each group. We perform a series of Monte Carlo simulations wherein films are randomly assigned to one of the three groups while each total group size remains constant. In this way, we are able to calculate the expected proportions of incoming and outgoing connections of each group if countries and languages of films were unimportant. If the randomized proportions are significantly different from the actual proportions, we must conclude that there are country and language biases.

Our analysis reveals that USA films have a disproportionate fraction of incoming and outgoing connections (Figure 4). This is a strong indicator of the USA and English-language biases present in the data set. The USA bias is further evident when we perform second-order analysis on the fractions of connections. We find that USA films receive percentages of the outgoing connections from other USA films and English non-USA films that are greater than the percentage of USA nodes in the network (Table 1). In addition, non-English non-USA films cite USA films nearly as often as they cite other non-English films (Table 1). Although this is clear evidence of bias for American films, it is not necessarily indicative of an American bias in IMDb's user base. IMDb is not an American-centric website, as it was originally founded in the United Kingdom (Needham, 2010). Moreover, the United States only accounts for 31% of IMDb's total traffic (Alexa, 2013c). (For comparison, the USA comprises 30% of Google's traffic and 30% of Apple's traffic; Alexa [2013a, 2013b].) More likely, the overrepresentation of USA films in the network reflects the pervasiveness of the American film industry around the world. For example, between 2007 and 2009, 23 of the 27 most-viewed films worldwide originated from the United States, and the other 4 were coproduced by U.S. companies (Acland, 2012). As such, more users are likely to identify citations to and from American films because they are the most-viewed films worldwide.

DOI: 10.1002/asi

TABLE 2. Characteristics of film-connection networks. Number of films and connections in the entire network, the giant component of the entire network, and the giant component of the network of films in the USA group only. Numbers in parentheses indicate the percentage of nodes or edges in the entire network.

	Entire network	Giant component	Giant component of USA group
Films	32,636	28,743 (88%)	15,425 (47%)
Connections	77,193	74,164 (96%)	42,794 (55%)

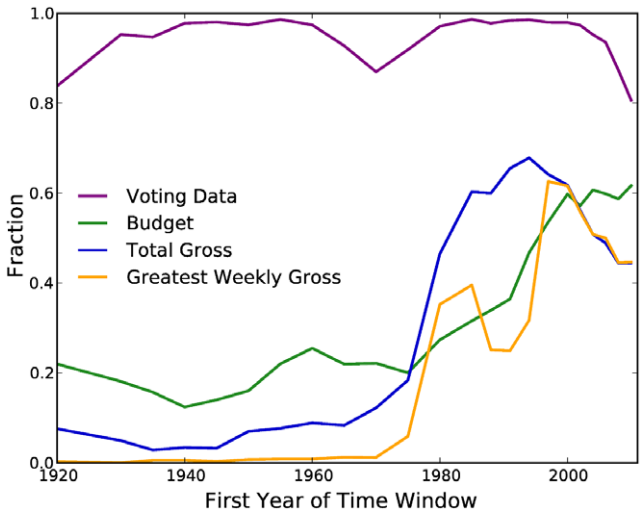

FIG. 5. Prevalence of reported data for USA films. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Due to the American-centric nature of both the financial information and the connections network as a whole, we chose to restrict our focus to films produced in the United States. This choice removes confounders caused by country of production from our analysis. Our new data set is the giant component of the connections network for USA films, which consists of 15,425 films and 42,794 connections (Table 2).

Missing Information

Not all IMDb entries report budget, total gross, or weekly gross information. The amount of missing data increases for older films, with "greatest weekly gross" being the most affected category of data (Figure 5). It appears that the weekly box office take may not have been a regularly reported statistic until the 1980s, when a sizable increase occurs in the reporting of greatest weekly gross. In fact, the weekly gross data found on Box Office Mojo—a website affiliated with IMDb—cover only the period after 1980 (Box Office Mojo, 2013). Due to the lack of reported weekly gross data prior to 1975, we choose to omit greatest weekly gross from our subsequent analysis.

Some IMDb entries also do not report a film's average user rating or number of user votes received. This reflects part of IMDb's method for calculating average user rating because the website does not post voting figures until a film has received a minimum of five user votes. Recent films are the most likely to lack voting data (Figure 5), which may be explained by users waiting to rate films until after viewing them.

Results

We proceed to examine the distributions of the values for three film statistics: production budget, total gross, and total number of user votes. We study the logarithm of these quantities because their values span several orders of magnitude. To minimize the effects of temporal bias, we look at the distributions for sets of films within 23 time windows spanning the period 1920 to 2011. We vary the length of the time windows to ensure that the number of films in each window is approximately constant. In addition, looking at financial values within narrow time windows mitigates the effects of GDP growth.

For each time window, we find the best-fitting Gaussian and double-Gaussian distribution parameters for the relevant data values. We then use bootstrapping to determine the statistical significance of the fits.

Because the double-Gaussian model is defined as a linear combination of two "single-" Gaussian models—thus having five parameters instead of two—the double-Gaussian model provides a better fit for the data under most circumstances. Therefore, to accept the double-Gaussian model as the true distribution, we must reject the Gaussian model as a possible fit and fail to reject the double-Gaussian model. If we fail to reject the Gaussian model, we assume that it is the correct description for the data. Using the Bonferroni correction, the threshold for rejecting a model is $p_B = 0.00217$ (Shaffer, 1995).

Among the statistics, the logarithm of total U.S. gross has the strongest evidence for the double-Gaussian distribution (Figure 6). We take the Gaussian model as its best representation prior to 1980. From 1980 on, we consider the double-Gaussian distribution as the most plausible model (Figure 7).

The double-log normal distribution of total U.S. gross was previously reported in 2010 by Pan and Sinha for films released between 1999 and 2008 (Pan & Sinha, 2010). In a 2013 paper, Chakrabarti and Sinha demonstrated that such bimodality can arise in a stochastic model where theaters independently decide which films to show (Chakrabarti & Sinha, 2013). We believe that the lower peak in the log total gross distribution includes big-budget films that flopped as well as "independent" films and "art" films that do not have the circulation afforded major studio releases. The appearance of the lower mode in the period 1980 to 1984 (Figure 7) corresponds to a time of rapid growth in the size and number of movie theaters in the United States, including small, independent film theaters. Between 1980 and 2000, the number of movie screens more than doubled as multiplexes

DOI: 10.1002/asi

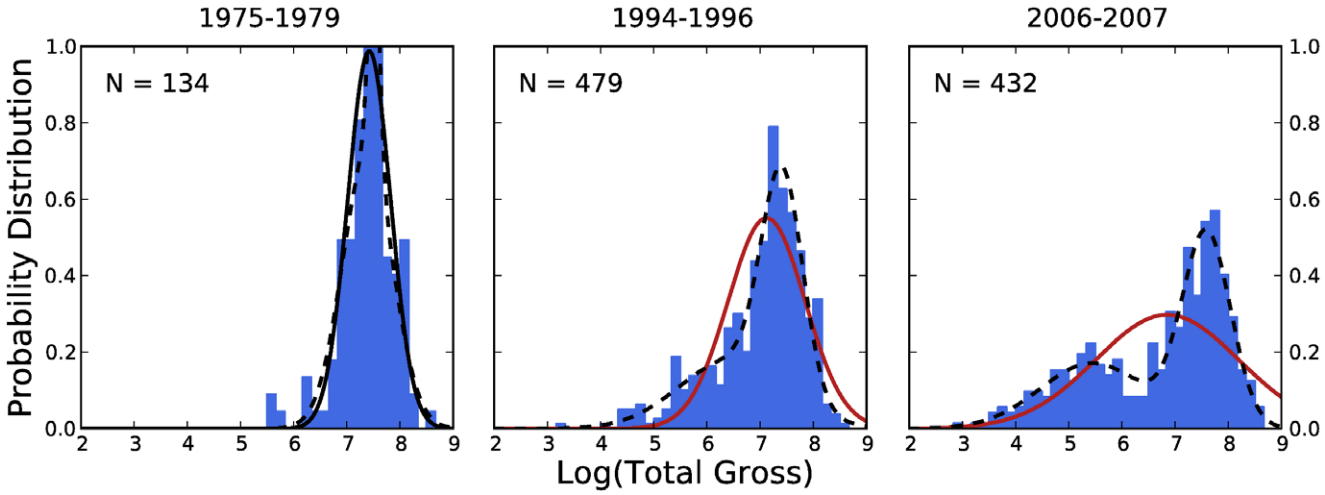

FIG. 6. Examples of distributions of log total gross. Distribution of the log of total U.S. gross in different time windows. The lines represent the best fits of a single-Gaussian distribution (solid line) and a mixture of two Gaussian distributions (dashed line) to the data. The color of the curve signifies whether we reject (red) or fail to reject (black) the distribution as a possible fit for the data. The number in the upper left corner is the total number of data points in the sample. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

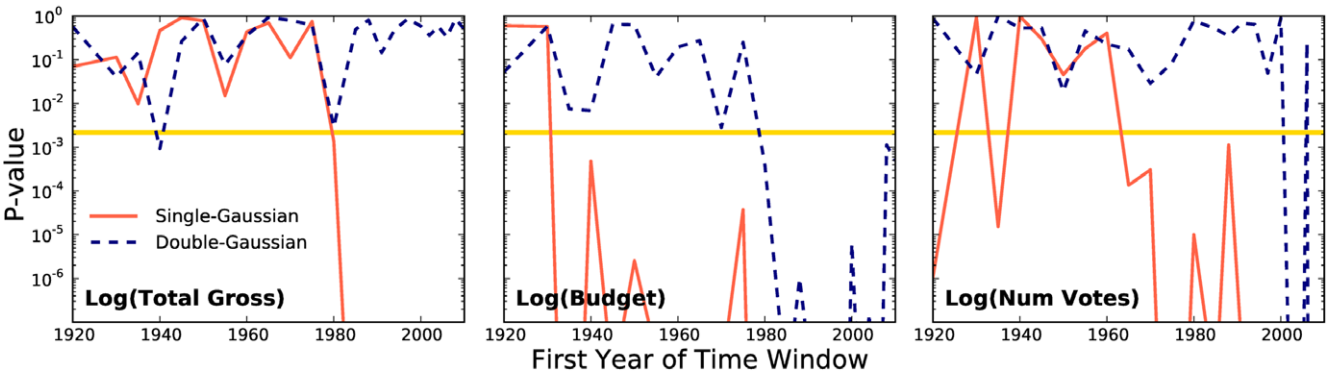

FIG. 7. *P* values of distribution fits over time. *P* values from bootstrapping analysis representing the goodness of fit of single-Gaussian and double-Gaussian distributions in specific time windows for the four considered statistics: the log of total gross (left), the log of film budget (center), and the log of number of user votes (right). The gold line represents the threshold *P* value—calculated according to the Bonferroni correction—below which we reject the distribution. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

replaced single-auditorium theaters (Noam, 2009). In addition, during the early 1980s, art theaters began to arise in smaller urban and suburban areas, whereas prior to that period, art theaters were only located in a select few major cities (McLane, 2002). This expansion of movie theaters enabled more low-budget films to be viewed by the general public and to have the opportunity for financial success.

For most of the considered time windows, the log of budget data exhibits a double-Gaussian distribution (Figure 8). This is the case for all time windows between 1935 and 1979 (Figure 7). After 1979, we reject both the single- and double-Gaussian distributions as potential fits (Figure 7). We suspect that the rejection of the unimodal and bimodal fits may be caused by the appearance of additional modes in the data beginning in 1980. This would place the emergence of new modes around the time of a reduction in cost of film-production equipment such as stereo-sound recorders (Enticknap, 2005). These new modes persist through the 1990s, as filmmaking switched from analog to digital.

The distributions for the log of total number of votes behave in a similar fashion. The data are initially (1920–1964) best represented by a Gaussian distribution followed by a period (1965–2001) where a mixture of two Gaussian distributions works well. After 2000, neither of the two proposed models is a good fit (Figure 7). The rejection of both proposed distributions after 2002 aligns with a marked rise in Internet usage and a large traffic increase for IMDb.

DOI: 10.1002/asi

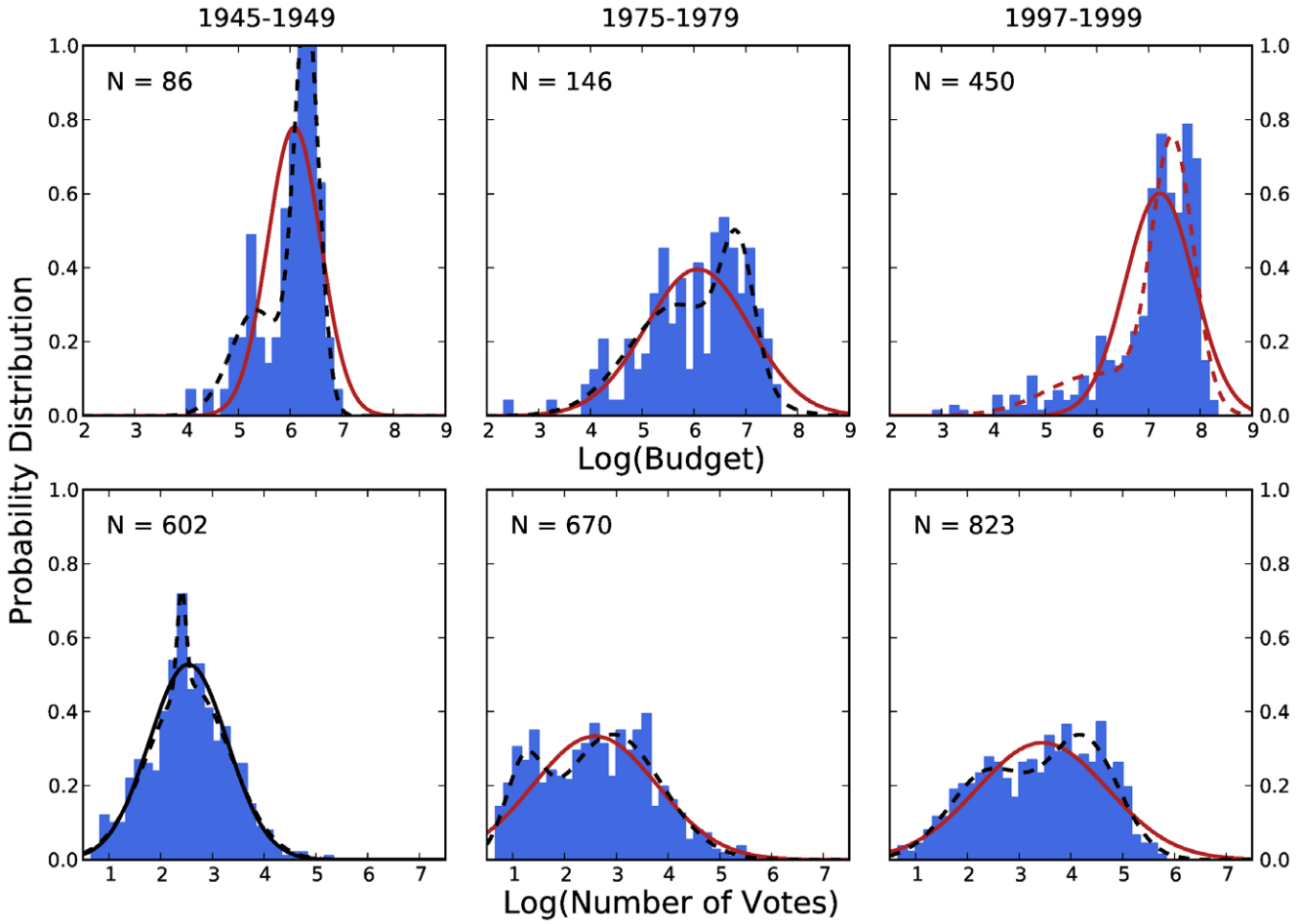

FIG. 8. Examples of distributions of log budget and log number of user votes. Distribution of the log of budget (upper row) and the log of number of votes (lower row) in different time windows. The lines represent the best fits of a single-Gaussian distribution (solid line) and a mixture of two Gaussian distributions (dashed line) to the data. The color of the curve signifies whether we reject (red) or fail to reject (black) the distribution as a possible fit for the data. The number in the upper left corner is the total number of data points in the sample. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Discussion

The number of votes received by a film is plausibly related to its quality and its financial characteristics. Thus, we next investigate a linear regression model for the log of number of votes.

$$v(k) = a_0 + a_1\hat{b}(k) + a_2\hat{g}_T(k) + a_3 r(k), \quad (1)$$

where $v(k)$ is the log number of user votes received by a film k, $\hat{b}(k)$ is the year-normalized log budget, $\hat{g}_T(k)$ is the year-normalized log total box office gross, and $r(k)$ is the average user rating. We normalize the monetary statistics by year to account for inflation, increases in ticket prices, and population growth. The procedure for calculating year-normalized log budget consists of subtracting a film's log budget value by the median of log budgets for all films released in the same calendar year,

$$\hat{b}(k) = b(k) - \tilde{b}_{y(k)} \quad (2)$$

where $b(k)$ is the actual log budget of film k, $\tilde{b}_y$ is the median of log budgets for films released in year y, and $y(k)$ is the year of release. Normalized total gross is computed in the same fashion as in Equation 2.

To properly estimate the parameter values for the model, we must account for the high prevalence of missing metadata. To do this, we use the Heckman correction method (Heckman, 1976, 1979) to adjust for selection bias caused by the absence of financial or voting data in approximately two thirds of films in the data set. For this method, we utilize a linear probit selection model to estimate the probability that a film in the data set reported the necessary information,

$$\Pr(\text{data reported}) = d_0 + d_1 y(k) + d_2 [i(k)]^{1/3} + d_3 [o(k)]^{1/3}, \quad (3)$$

where $i(k)$ is the in-degree of film k, and $o(k)$ is its out-degree. This conditional probability is then applied as a correction term to Equation 1.

DOI: 10.1002/asi

TABLE 3. Correlations among year-normalized financial data, user rating, and user votes. Results of correlation analyses comparing the log of number of user votes to the log of normalized budget, the log of normalized total gross, and the average user rating, using the Heckman correction method to account for missing data. Each column represents a different implementation of the linear model in Equation 1. Due to limited space, we do not display results for two of the possible models: The model using budget and total gross as independent variables, and the model using total gross and user rating as independent variables. We omit the budget and gross model because it is outperformed by the model using only budget (adjusted R^2 of 66.59 vs. 73.97%) because budget and gross are highly correlated. We omit the gross and rating model because it is outperformed by the model using budget and user rating (68.31 vs. 75.50%). Except for the coefficient for year and the intercept in the user rating-only model, all coefficient values are significant with $p < 0.001$. These results suggest that budget alone explains much of the variation in number of votes.

	$r$only		$\hat{g}_T$ only		$\hat{b}$ only		$\hat{b}$ and r		All	
Model	Coeff.	SE.	Coeff.	SE.	Coeff.	SE.	Coeff.	SE.	Coeff.	SE
Log of user votes										
Intercept	3.16	0.05	4.67	0.02	4.85	0.03	4.19	0.04	3.80	0.04
Log norm. budget $\hat{b}$	—	—	—	—	0.602	0.006	0.586	0.006	0.268	0.01
Log norm. total gross $\hat{g}_T$	—	—	0.354	0.006	—	—	—	—	0.191	0.008
User rating r	0.080	0.005	—	—	—	—	0.102	0.006	0.156	0.006
Probit selection										
Intercept	0.348	1.5	−80	2	−46	1	−46	1	−76	2
Year y	-5×10^{-5}	8×10^{-4}	0.040	8×10^{-4}	0.023	6×10^{-4}	0.023	6×10^{-4}	0.037	9×10^{-4}
Cube root in-degree $i^{1/3}$	1.19	0.07	0.916	0.02	0.713	0.02	0.713	0.02	0.884	0.02
Cube root out-degree $o^{1/3}$	0.915	0.06	0.354	0.02	0.250	0.02	0.250	0.02	0.317	0.02
Inverse mills ratio	−6.47	0.2	−0.81	0.02	−1.16	0.02	−1.12	0.02	−0.61	0.01
Total no. of films	15,425		15,425		15,425		15,425		15,425	
Films with observed data	14,577		5,307		5,331		5,331		3,430	
Films with censored data	848		10,118		10,094		10,094		11,995	
Adjusted R^2 %	**26.02**		**63.74**		**73.97**		**75.50**		**72.83**	

From the correction analysis, we find that both year-normalized financial statistics correlate strongly with total number of user votes (Table 3). The stronger correlation exists between the log of number of user votes and the log of normalized film budgets. When both business statistics are used in a linear model for number of votes, the correlation is not as strong as when the log of budget is used alone. The strongest correlation, however, exists for the use of log of normalized budget in conjunction with average user rating (Table 3). Interestingly, when we replace log of user votes with user rating as the dependent variable, we find no correlation with either financial quantity.

Our result suggests that the number of user votes is an indicator of a film's prominence. After all, a person is more likely to enter a rating for a film if he or she has viewed it, regardless of whether he or she found the film to be good or bad. However, prominence is not necessarily tied to box office success, as many films become notable for other reasons such as major award nominations. Films also can become notable for especially poor performances at the box office (e.g., 1995's *Cutthroat Island*, which cost $98 million to produce and only grossed $10 million). In addition, the production budget for a film—also known as the "negative cost"—has been found to be strongly correlated with a film's advertising cost (Prag & Casavant, 1994). Hence, it is understandable that the total number of user votes correlates more strongly with budget than with box office gross, as the former is directly related to the amount spent on a film's promotion, which increases its prominence. Moreover, we find that film quality—in the form of average user rating—does not appreciably account for prominence when applied in conjunction with budget in the linear model. Therefore, budget is overwhelmingly the most relevant factor in determining a film's ultimate prominence. To make a film more notable, Hollywood does not need to spend more money on making it better; Hollywood just needs to spend more money.

Acknowledgments

We thank Irmak Sirer, Andrea Lancichinetti, Dr. Alan Wasserman, Adam Hockenberry, Julia Poncela, João Moreira, and the members of the Amaral Lab for their comments and suggestions. We thank Penny Dash for editorial assistance.

References

Acland, C.R. (2012). From international blockbusters to national hits: Analysis of the 2010 UIS survey on feature film statistics. UNESCO Institute for Statistics Information Bulletin no. 8. Retrieved from http://www.uis.unesco.org/FactSheets/Documents/ib8-analysis-cinema-production-2012-en2.pdf

Alexa (2013a). Apple.com site info. Retrieved from http://www.alexa.com/siteinfo/apple.com

Alexa (2013b). Google.com site info. Retrieved from http://www.alexa.com/siteinfo/google.com

Alexa (2013c). Imdb.com site info. Retrieved from http://www.alexa.com/siteinfo/imdb.com

Amaral, L.A.N., Scala, A., Barthelemy, M., & Stanley, H.E. (2000). Classes of small-world networks. Proceedings of the National Academy of Sciences, USA, 97(21), 11149–11152.

Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proceedings of the National Academy of Sciences, USA, 106(51), 21544–21549.

DOI: 10.1002/asi

Box Office Mojo. (2013). About movie box office tracking and terms. Retrieved from http://www.boxofficemojo.com/about/boxoffice.htm.
Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. Reviews of Modern Physics, 81(2), 591–646.
Chakrabarti, A.S., & Sinha, S. (2013). Self-organized coordination in collective response of non-interacting agents: Emergence of bimodality in box-office success. Retrieved from http://arxiv.org/abs/1312.1474
Ebel, H., Mielsch, L.-I., & Bornholdt, S. (2002). Scale-free topology of e-mail networks. Physical Review E, 66(3), 035103.
Enticknap, L. (2005). Moving image technology: From zoetrope to digital. London, U.K.: Wallflower Press.
Fairbanks, D. (Producer), & Walsh, R. (Director). (1924). The thief of Bagdad [Motion picture]. United States: United Artists.
Golder, S.A., & Macy, M.W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science, 333(6051), 1878–1881.
Gonçalves, B., & Ramasco, J. (2008). Human dynamics revealed through Web analytics. Physical Review E, 78(2), 026123.
Grujić, J. (2008). Movies recommendation networks as bipartite graphs. In M. Bubak, G.D. Van Albada, J. Dongarra, & P.M.A. Sloot (Eds.), Computational Science–International Conference on Computational Science 2008 (ICCS 2008, pp. 576–583). Krakow, Poland: Springer-Verlag Berlin.
Harlin, R. (Producer, Director), Mark, L. (Producer), Michaels, J.B. (Producer), & Gorman, J. (Producer). (1995). Cutthroat island [Motion picture]. United States: Metro-Goldwyn-Mayer.
Heckman, J.J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement, 5(4), 475–492.
Heckman, J.J. (1979). Sample selection bias as a specification error. Econometrica: Journal of the Econometric Society, 47(1), 153–161.
Herr, B., Ke, W., Hardy, E., & Börner, K. (2007). Movies and actors: Mapping the Internet movie database. In E. Banissi, R. Burkhard, G. Grinstein, U. Cvek, M. Trutschl, L. Stuart, & A. Ursyn (Eds.), Proceedings of the 11th International Conference on Information Visualization (pp. 465–469). Zurich, Switzerland: IEEE Computer Society.
Hertzberg, M. (Producer), & Brooks, M. (Director). (1974). Blazing saddles [Motion picture]. United States: Warner Bros.
Johansen, A. (2001). Response time of internauts. Physica A, 296(3–4), 539–546.
Malmgren, R.D., Stouffer, D.B., Motter, A.E., & Amaral, L.A.N. (2008). A Poissonian explanation for heavy tails in e-mail communication. Proceedings of the National Academy of Sciences, USA, 105(47), 18153–18158.
Malmgren, R.D., Stouffer, D.B., Campanharo, A.S.L.O., & Amaral, L.A.N. (2009). On universality in human correspondence activity. Science, 325(5948), 1696–1700.
McLane, B. (2002). Domestic theatrical and semi-theatrical distribution and exhibition of American independent feature films: A survey in 1983. In G.A. Waller (Ed.), Moviegoing in America (pp. 265–267). Oxford, United Kingdom: Blackwell.
Needham, C. (2010). IMDb 20th anniversary: A letter from our founder. Retrieved from http://www.imdb.com/features/anniversary/2010/letter
Noam, E.M. (2009). Media ownership and concentration in America. New York, NY: Oxford University Press.
Pan, R.K., & Sinha, S. (2010). The statistical laws of popularity: Universal properties of the box-office dynamics of motion pictures. New Journal of Physics, 12(11), 115004.
Prag, J., & Casavant, J. (1994). An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry. Journal of Cultural Economics, 18(3), 217–235.
Radicchi, F. (2009). Human activity in the web. Physical Review E, 80(2), 026118.
Reiner, R. (Producer, Director), Scheinman, A. (Producer), & Ephron, N. (Producer). (1989). When Harry met Sally . . . [Motion picture]. United States: Columbia Pictures.
Shaffer, J.P. (1995). Multiple hypothesis testing. Annual Review of Psychology, 46(1), 561–584.
Spielberg, S. (Producer, Director), & Kennedy, K. (Producer). (1982). E.T. the extra-terrestrial [Motion picture]. United States: Universal Pictures.
Sreenivasan, S. (2013). Quantitative analysis of the evolution of novelty in cinema through crowdsourced keywords. Scientific Reports, 3, 2758.
Wallis, H.B. (Producer), & Curtiz, M. (Director). (1942). Casablanca [Motion picture]. United States: Warner Bros.
Wanger, W. (Producer), & Ford, J. (Director). (1939). Stagecoach [Motion picture]. United States: United Artists.
Watts, D.J., & Strogatz, S.H. (1998). Collective dynamics of "small-world" networks. Nature, 393(6684), 440–442.
Wikimedia Foundation. (2011). Wikipedia editors study: Results from The Editor Survey, April 2011. Retrieved from http://upload.wikimedia.org/wikipedia/commons/7/76/Editor_Survey_Report___April_2011.pdf
Woolley, A.W., Chabris, C.F., Pentland, A., Hashmi, N., & Malone, T.W. (2010). Evidence for a collective intelligence factor in the performance of human groups. Science, 330(6004), 686–688.
Wuchty, S., Jones, B.F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. Science, 316(5827), 1036–1039.

DOI: 10.1002/asi