Supporting Information

Wasserman et al. 10.1073/pnas.1412198112

SI Text

Network. To control for biases in data reporting in IMDb (1), we choose to consider only the network of connections between films made in the United States. Furthermore, we only select connections designated on IMDb as references, spoofs, or features. We also limit our analysis to films released in 2011 or earlier. We obtain the IMDb film connections information—as well as information on country of production, primary language, and production companies—from plain text data files provided through ftp sites (2). We use a Python (version 2.7.8) program developed in-house to parse the relevant information from this file.

After entering the valid connections, we construct a network where each connection is a directed edge and each film is a node (Fig. 1). A link from movie A to movie B signifies that movie A cites movie B. To ensure proper maintenance of the timeline, we only include a connection from A to B if A was released in a later calendar year than B. As such, the network contains no links that are "forward" in time and no links between two films released in the same calendar year. Therefore, the resulting network is acyclic. We then take the largest weakly connected component of this network, known as the giant component, for our analysis.

Data. After the network is constructed, we count the number of times each film is cited (in-degree) and the number of citations each film makes (out-degree). We compute the PageRank value for each film in the network using the NetworkX Python package (version 1.8.1). The formula for the time lag of a citation is as follows:

$$t = y(k_{\text{out}}) - y(k_{\text{in}}), \qquad \text{[S1]}$$

where $y(k)$ is the year of release of film k, and k_{out} and k_{in} are the films on the outgoing and incoming sides of an edge, respectively. After calculating the time lag for every edge in the network, we count the number of citations with time lag of at least 25 y that each film receives. This is the long-gap citation count.

We collect data on IMDb average user ratings and total numbers of votes for each film in the network through provided text files (2). Data on box office information and genre are also obtained through these files. We use Python programs developed in-house for parsing these files. The IMDb ID numbers for each film, which are necessary for accessing a film's page on the IMDb website (www.imdb.com), are obtained through an in-house web scraping Python program using the BeautifulSoup package (version 4.3.2). We use this package in all our web-scraping processes.

We scrape Metacritic scores for films from web pages on the Metacritic website (www.metacritic.com). Each Metacritic web page is accessed via a film's "critic reviews" page on the IMDb website, which contains a direct link to Metacritic if an aggregate review score exists for that film. We scrape Roger Ebert ratings for films from pages on Ebert's official site (www.rogerebert.com). Each page on Ebert's site is accessed through a film's "external reviews" page on IMDb, which consists of user-added links to reviews of films on external websites. If Roger Ebert reviewed a film, a link to his review generally appears first on this page. We manually compile the list of films present in the National Film Registry (NFR) as the limited number makes this option possible (3).

Distribution Modeling. To generate null models for the distribution of time lags in the film connections network, we create Markov chain Monte Carlo simulations wherein the network undergoes random rewiring (4–7) (Fig. S2). In each step of a simulation, two edges are selected at random from the network of films as candidates for rewiring. If the candidate connections "overlap"—that is, if at least one of the films of edge E was released in a calendar year in-between the years of release of the two films of edge F, noninclusive—then a swapping of connection nodes occurs. The "swapping" process consists of removing the chosen edges E and F from the network and replacing them with two new edges G and H, where edge G connects the outgoing film of edge E to the incoming film of edge F, and edge H connects the outgoing film of edge F to the incoming film of edge E. By allowing swapping between overlapping edges, we ensure that no back-in-time links are created. We forbid swapping if one of the edges created as a result of swapping already exists in the network. This process allows for random redistribution of edges while maintaining the in- and out-degrees of all of the nodes.

We use the simulation to generate the base null model—where the two randomly chosen edges are always swapped when it is legal to do so—as well as a null model with a bias toward shorter-length citations. In these latter simulations, a legal pair of randomly chosen edges undergoes swapping with probability q:

$$q = e^{[\min(t_1,t_2)-\min(s_1,s_2)]/40}, \qquad \text{[S2]}$$

where t_1 and t_2 are the time lags of the two chosen edges and s_1 and s_2 are the time lags of the two edges if they were to be swapped. More specifically, if $t_1 = y_1 - z_1$ and $t_2 = y_2 - z_2$, where y_i and z_i are the years of the films connected by edge i, then $s_1 = y_1 - z_2$ and $s_2 = y_2 - z_1$.

In each run of a simulation, $20n_e$ iterations are performed—where n_e is the number of edges in the network. In total, we run 400 simulations, 200 with the base simulation and 200 with the biased simulation. We use Python programs developed in-house to run all rewiring simulations.

In addition, we use a theoretical formula for the time lag distribution of the unbiased null model, given nodes with specific in-degrees, out-degrees, and years of release:

$$\mathbb{E}(L_t) = \sum_{y \in \mathcal{Y}} \mathbb{E}\left(c_{y,y-t}\right), \qquad \text{[S3]}$$

where L_t is the number of links with time lag t in the null model, $\mathcal{Y}$ is the set of all years of release for films in the network, and $c_{y,z}$ is the number of links between films released in year y and films released in year z $(y > z)$. The expected value of $c_{y,z}$ is determined by the following formula:

$$\mathbb{E}\left(c_{y,z}\right) = \frac{o_y i_z}{\sum_{\substack{j \in \mathcal{Y} \\ j<y}} (i_j - o_j)} \prod_{k=z+1}^{y-1} \left(1 - \frac{o_k}{\sum_{\substack{j \in \mathcal{Y} \\ j<k}} (i_j - o_j)}\right), \qquad \text{[S4]}$$

where i_y and o_y are the sum totals of in-citations and out-citations, respectively, for films released in year y. We adapt this equation from Karrer and Newman's formula for the expected number of edges between vertices in a directed acyclic graph with a fixed degree sequence (8).

Linear Regression. We narrow our focus to seven metrics: Roger Ebert rating, Metacritic score, IMDb average user rating, number of IMDb votes, total citation count, PageRank score, and long-gap

citation count. We calculate the adjusted R^2 values of linear regressions between each pair of considered measures using the statsmodels Python package (version 0.5.0). In our linear regression models, we use the base-10 logarithm of IMDb votes and PageRank score and the cube root of citations and long-gap citations. We opt to use cube root rather than log for citations and long-gap citations because many films have 0 values for these metrics, and positive values extend over several orders of magnitude. All regressions we perform in this paper apply these functions to these metrics.

Probit Regression. We perform probit regressions of the following form:

$$inNFR \sim SigMetric, \quad \textbf{[S5]}$$

where *inNFR* is the categorical variable representing whether or not a film is in the NFR (1 if it is in the NFR and 0 if it is not) and *SigMetric* is one of the seven metrics. For metrics with missing data—which are the expert-based metrics and the IMDb voting statistics—we apply the Heckman correction method (9, 10) to the probit regression, using R (version 3.0.2) and the sampleSelection package (version 1.0-2) (11). For metrics without missing data, we perform the regression with the statsmodels package in Python. We use probit instead of logit for this analysis because the sampleSelection package can only apply the Heckman correction for binary outcomes using probit.

When we apply the Heckman correction method, we use year of release and film genre as the dependent variables in the selection model equation. We note that genre is actually a set of 24 binary variables representing the 24 categorical film genres listed on IMDb. Films are not limited to being classified as one genre, and 11,661 of films in the network (or 75.6%) are categorized under two or more genres.

For all of these models apart from the long-gap citations model, we perform the regression on the subset of films released on or before 2003, as only films released on or before that year were eligible for nomination to the NFR in 2013 (3). For the long-gap citations model, we perform the regression on the subset of films made on or before 1986. The justification for the different subsets is that all films released after 1986 in our dataset have zero long-gap citations. (Our dataset only includes films released up to 2011, and the latest year that can possibly have a nonzero number of citations with a 25-or-more-year time lag is 1986.)

From the probit regression models, we obtain estimated probabilities for each film used to create the model. From these estimated probabilities, we assign a predicted value of 0 or 1 to each observation (0 if the probability is below 0.5, and 1 if the probability is greater than or equal to 0.5). We use the actual and predicted values to construct the classification table. We use the classification table to compute the balanced accuracy:

$$\begin{aligned}\text{Balanced Accuracy} &= \frac{\text{Sensitivity} + \text{Specificity}}{2} \\ &= \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right),\end{aligned} \quad \textbf{[S6]}$$

where TP, TN, FP, and FN are true positives, true negatives, false positives, and false negatives, respectively. We use the balanced accuracy instead of the true accuracy because the latter is strongly affected by the imbalance toward films not in the NFR versus those that are.

We also use the estimated probabilities to determine the receiver operating characteristic (ROC) curve (12). We calculate the area under the ROC curve (AUC) with the scikit-learn Python package (version 0.14.1). Finally, we use our probit regression results to calculate the pseudo-R^2 using Tjur's equation (13).

We obtain SD values for the balanced accuracy, AUC, and pseudo-R^2 of each metric using bootstrapping with 1,000 random samples. We use programs developed in-house in either Python or R to conduct all of the bootstrapping we do for this project.

Additionally, we repeat the same analysis detailed in this section, only instead of accounting for missing data, we ignore it and perform probit regression on the reported values (Table S1). This allows us to clearly present the effect of missing data and the Heckman correction.

Random Forest Classification. We perform Random Forest (RF) classification (14) using R and the randomForest package (version 4.6–10) (15). We conduct RF classification once with all seven aforementioned metrics as predictor variables, and another time with all metrics apart from long-gap citation count. In both cases, we use presence in the NFR as the binary response. We perform cross-validation by conducting 100 iterations of RF classification with each iteration using 80% of the data points, chosen randomly without replacement. We use the subset of films that were made in 1999 or earlier and have reported data for all seven metrics in our RF classification. This subset consists of 766 films. We conduct each classification iteration using 1,000 classification trees. From the cross-validated RF classification results, we obtain the mean and SD of variable importance—also known as the permutation importance—for each predictor.

Multivariate Regression. We perform two probit regressions using multiple independent variables. The first uses all metrics apart from long-gap citations as independent variables, whereas the second includes long-gap citations. In both regressions, the dependent variable is presence in the NFR. Also, both regressions are performed on the same subset of films used in RF classification. We evaluate the fit of the regression models by calculating the pseudo-R^2 with McFadden's equation (16). We perform these regressions with the statsmodels Python package.

We also repeat this same analysis but with logit instead of probit to demonstrate the minimal differences between the regression models (Table S2).

Citation Description Analysis. The brief notes that accompany some film connections on IMDb are not provided in the aforementioned plain text files, which we originally used to construct the connections network. Instead, we obtain these descriptions by scraping them from the actual IMDb movie connections pages. For each citation in the network, we check the cited film's connections page to see first whether the citing film is listed, and second whether a description is included with that citation. If a citing film is listed twice on the page and each listing has a description, then both descriptions are scraped. As with the initial construction of the network, we only scrape a description if the citation is classified as a reference, spoof, or feature.

After obtaining the citation descriptions, we proceed with two methods of analysis. In the first method, we take a small subset of highly cited films and, by hand, classify all of the annotations based on what they are citing. The subset of films we consider is the bottom 15 films from Table S3 (i.e., from *Bride of Frankenstein* to *Dirty Harry*). We classify annotated citations as "general" if the annotations merely refer to a film's title, title character, or plot, or if the citation is to numerous clips of the film. If an annotation is not general, then we classify the citation according to the part of the film to which it pertains, such as a specific scene, quotation, character, setting, or song. Two people independently classified the annotated citations for these films. The two people differed by no more than two citations in any classification for any film. The results of this manual classification are shown in Table S4.

In the second method, we use the token_set_ratio function from the fuzzywuzzy Python package (version 0.3.2) to perform

comparisons between citation descriptions. Initially, we clean all of the descriptions by removing all punctuation—apart from hyphens and apostrophes in-between letters—and converting all alphabet characters to lowercase. For each film with a minimum of 20 annotated citations, we compare every pair of descriptions using the token_set_ratio function, which returns an integer value indicating the similarity of two strings, with 0 being the least similar and 100 being the most similar. Thus, for a film with c annotated citations, we obtain $\binom{c}{2}$ similarity values. Taking the average of all of the similarity values for a film gives us the "mean similarity" for that film's citation descriptions.

To compensate for differing numbers of descriptions and varying lengths of strings, we perform bootstrapping wherein all of the words in all of the descriptions for a specific film are randomly redistributed while keeping the number of words in each description constant. We then perform the aforementioned process for computing the mean similarity on the jumbled citation descriptions. We perform 500 randomization iterations for each film with a minimum of 20 annotated citations. We then obtain a mean and SD for all of the randomized mean similarities for a film, as well as a Z score for the mean similarity of the actual descriptions. We perform linear regressions comparing the Z scores to the results of manual classification.

1. Wasserman M, et al. (2014) Correlations between user voting data, budget, and box office for films in the Internet Movie Database. *J Am Soc Inf Sci Technol*, 10.1002/asi.23213.
2. Internet Movie Database (2012) Alternative Interfaces. Available at www.imdb.com/interfaces. Accessed October 26, 2012.
3. Library of Congress (2014) National Film Registry. Available at www.loc.gov/film/filmnfr.html. Accessed April 11, 2014.
4. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. *Science* 296(5569):910–913.
5. Milo R, et al. (2002) Network motifs: Simple building blocks of complex networks. *Science* 298(5594):824–827.
6. Milo R, Kashtan N, Itzkovitz S, Newman MEJ, Alon U (2004) On the uniform generation of random graphs with prescribed degree sequences. arXiv:cond-mat/0312028.
7. Carstens C (2013) Motifs in directed acyclic networks. *2013 International Conference on Signal-Image Technology and Internet-Based Systems* (IEEE Computer Society, Los Alamitos, CA), pp 605–611.
8. Karrer B, Newman MEJ (2009) Random graph models for directed acyclic networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 80(4 Pt 2):046110.
9. Heckman JJ (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Ann Econ Soc Meas* 5(4):475–492.
10. Heckman JJ (1979) Sample selection bias as a specification error. *Econometrica* 47(1):153–161.
11. Toomet O, Henningsen A (2008) Sample selection models in R: Package sampleSelection. *J Stat Softw* 27(7):1–23.
12. Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clin Chem* 39(4):561–577.
13. Tjur T (2009) Coefficients of determination in logistic regression models—a new proposal: The coefficient of discrimination. *Am Stat* 63(4):366–372.
14. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32.
15. Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News* 2(3):18–22.
16. McFadden D (1974) *Conditional Logit Analysis of Qualitative Choice Behavior. Frontiers in Econometrics*, ed Zarembka P (Academic, New York), pp 105–142.

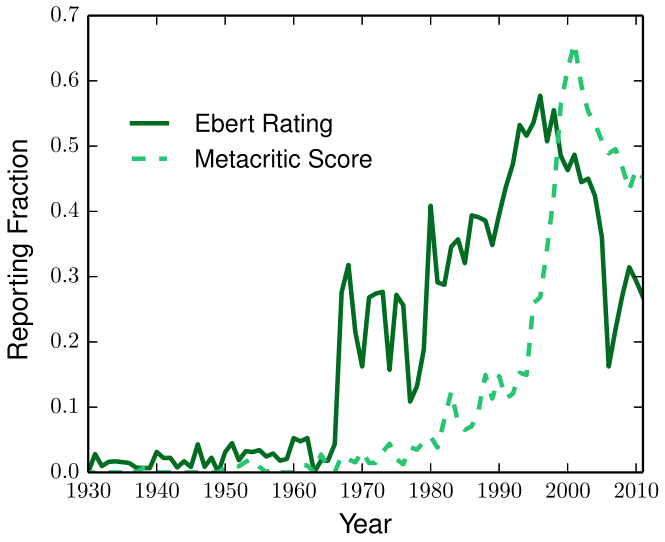

Fig. S1. Fraction of reported critic data values by year.

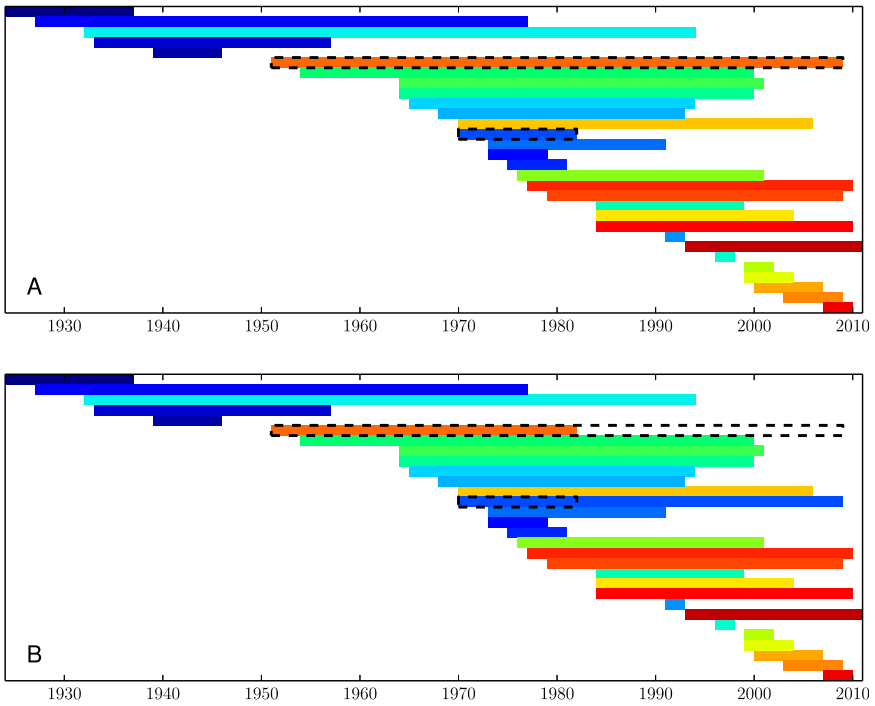

Fig. S2. One step of rewiring simulation. The diagram depicts a randomly chosen sample of edges from the directed network of film connections. Each bar represents an edge in the sample. The right end of each bar indicates the year of release for the citing film. The left end of each bar indicates the year of release for the cited film. (*A*) Two edges are selected at random as candidates for swapping. (*B*) If the two chosen edges overlap, they are removed from the network and replaced with two new edges that connect the outgoing film of one original edge to the incoming film of the other original edge. The black dotted lines represent the originally chosen candidate edges, now removed from the network.

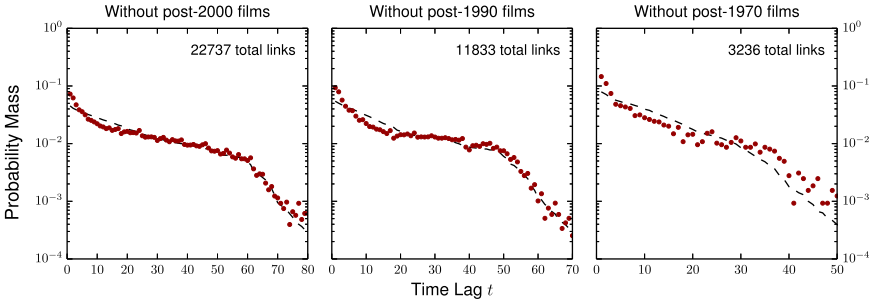

Fig. S3. Distribution of time lag with exclusions. Probability mass function of the time lag of connections in the film connections network, discounting all films made after 2000, after 1990, and after 1970. Brown points represent the actual distributions. Dashed black lines represent the unbiased null model distribution, calculated with Eq. **S3**.

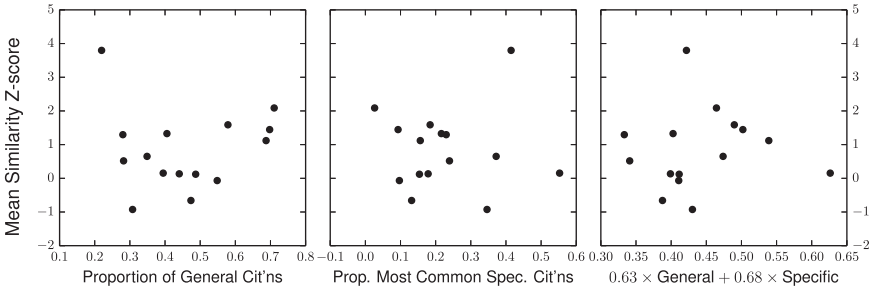

Fig. S4. Mean similarity *Z* score versus annotated citation classifications. Scatterplots comparing the mean similarity *Z* scores of citation annotations to the fractions of annotations under certain classifications for 15 highly cited films (See Table S4). The fractions considered are the proportion of general citations (*Left*), the proportion of most common specific citations (*Center*), and the best-fitting linear combination of the two proportions obtained through ordinary least-squares regression (*Right*). No adjusted R^2 value is positive.

Table S1. Binary regression results for several estimators of significance, ignoring missing data

Metric*	*N*	Fraction in NFR	Balanced accuracy[†]	AUC[‡]	pR^{2}[§]
Ebert rating[¶]	2,980	0.061	0.5 (0.)	0.87 (0.01)	0.16 (0.02)
Metacritic score[¶]	1,652	0.045	**0.61** (0.04)	**0.93** (0.01)	**0.27** (0.04)
IMDb average rating[¶]	11,805	0.039	0.502 (0.003)	**0.88** (0.01)	0.13 (0.01)
IMDb votes[¶]	11,805	0.039	0.5 (0.)	0.76 (0.01)	0.039 (0.005)
Total citations[¶]	12,339	0.037	0.57 (0.01)	0.86 (0.01)	0.19 (0.02)
PageRank[¶]	12,339	0.037	0.57 (0.01)	0.85 (0.01)	0.19 (0.02)
Long-gap citations[#]	8,011	0.054	**0.61** (0.01)	0.88 (0.01)	**0.26** (0.02)

*SDs in parentheses. Top two values for each performance category in bold.
[†]Obtained from classification table analysis with 0.5 as the threshold.
[‡]Area under the receiver operating characteristic (ROC) curve (12).
[§]Tjur's pseudo-R^{2} (13).
[¶]Regression performed on films released on or before 2003.
[#]Regression performed on films released on or before 1986.

Table S2. Contributions of several estimators of significance in multivariate logit regression

Model	pR^{2}*	ΔpR^{2}
Metacritic + IMDb rating +		
IMDb votes + total citations	0.6066	—
– Total citations	0.4924	−0.1142
– Metacritic	0.5382	−0.0684
– IMDb votes	0.5439	−0.0627
– IMDb rating	0.5538	−0.0528
Metacritic + IMDb rating +		
Long-gap citations	0.6260	—
– Long-gap citations	0.4866	−0.1394
– Metacritic	0.5547	−0.0713
– IMDb rating	0.5849	−0.0411

*McFadden's pseudo-R^{2} (16).

Table S3. Films with most long-gap citations

Title	Year	LGC*	NFR year†
The Wizard of Oz	1939	565	1989
Star Wars	1977	297	1989
Psycho	1960	241	1992
Casablanca	1942	212	1989
Gone with the Wind	1939	198	1989
King Kong	1933	191	1991
Frankenstein	1931	170	1991
The Godfather	1972	162	1990
Citizen Kane	1941	143	1989
2001: A Space Odyssey	1968	143	1991
Jaws	1975	129	2001
Night of the Living Dead	1968	122	1999
It's a Wonderful Life	1946	109	1990
The Graduate	1967	97	1996
Vertigo	1958	92	1989
Snow White and the Seven Dwarfs	1937	91	1989
Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb	1964	91	1989
Dracula	1931	90	2000
The Maltese Falcon	1941	80	1989
Bambi	1942	79	2011
The Exorcist	1973	78	2010
Taxi Driver	1976	71	1994
Sunset Blvd.	1950	70	1989
Planet of the Apes	1968	69	2001
Deliverance	1972	66	2008
The Sound of Music	1965	61	2001
Bride of Frankenstein	1935	58	1998
Singin' in the Rain	1952	57	1989
Apocalypse Now	1979	57	2000
The Texas Chain Saw Massacre	1974	57	
Rebel Without a Cause	1955	57	1990
Star Wars: Episode V—The Empire Strikes Back	1980	56	2010
North by Northwest	1959	54	1995
Rear Window	1954	54	1997
Mary Poppins	1964	54	2013
Pinocchio	1940	53	1994
Willy Wonka & the Chocolate Factory	1971	52	
The Seven Year Itch	1955	51	
Rosemary's Baby	1968	51	
West Side Story	1961	51	1997
Dirty Harry	1971	51	2012

*Long-gap citation count.
†Year inducted into the NFR (3).

Table S4. Classification of citation descriptions

	Annotated citations	General citations		Most common specific citations			General + specific
Title*	*N*	*N*	%	Description	*N*	%	%
Bride of Frankenstein	38	27	**71**	Multiple	1	3	74
Singin' in the Rain	43	15	35	Title scene/song	16	**37**	72
Apocalypse Now	100	28	28	"Smell of napalm"	23	23	51
The Texas Chain Saw Massacre	84	37	44	Leatherface	15	18	62
Rebel Without a Cause	38	22	58	Jim Stark	7	18	76
Star Wars: The Empire Strikes Back	117	33	28	Yoda	28	24	52
North by Northwest	26	8	31	Crop duster scene	9	35	65
Rear Window	37	15	41	Peeping scenes	8	22	62
Mary Poppins	43	30	**70**	Flying umbrella	4	9	**79**
Pinocchio	39	19	49	Jiminy Cricket	6	15	64
Willy Wonka & the Chocolate Factory	41	9	22	Oompa Loompas	17	**41**	63
The Seven Year Itch	38	15	39	Dress blowing scene	21	**55**	**95**
Rosemary's Baby	31	17	55	Apartment building	3	10	65
West Side Story	38	18	47	"I feel pretty"	5	13	61
Dirty Harry	64	44	**69**	"Do you feel lucky?"	10	16	**84**

*Three largest values in each percentage column are shown in bold.