Cross-evaluation of metrics to estimate the significance of creative works

Max Wasserman[a], **Xiao Han T. Zeng**[b], **and Luís A. Nunes Amaral**[b,c,d,e,1]

Departments of [a]Engineering Sciences and Applied Mathematics, [b]Chemical and Biological Engineering, and [c]Physics and Astronomy, [d]Howard Hughes Medical Institute, and [e]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved December 1, 2014 (received for review June 27, 2014)

In a world overflowing with creative works, it is useful to be able to filter out the unimportant works so that the significant ones can be identified and thereby absorbed. An automated method could provide an objective approach for evaluating the significance of works on a universal scale. However, there have been few attempts at creating such a measure, and there are few "ground truths" for validating the effectiveness of potential metrics for significance. For movies, the US Library of Congress's National Film Registry (NFR) contains American films that are "culturally, historically, or aesthetically significant" as chosen through a careful evaluation and deliberation process. By analyzing a network of citations between 15,425 United States-produced films procured from the Internet Movie Database (IMDb), we obtain several automated metrics for significance. The best of these metrics is able to indicate a film's presence in the NFR at least as well or better than metrics based on aggregated expert opinions or large population surveys. Importantly, automated metrics can easily be applied to older films for which no other rating may be available. Our results may have implications for the evaluation of other creative works such as scientific research.

data science | complex networks | citations | films | IMDb

For many types of creative works—including films, novels, plays, poems, paintings, and scientific research—there are important efforts for identifying which creations are of the highest quality and to honor their creators, including the Oscars, the Pulitzer Prize, and the Nobel. Unfortunately, these distinctions recognize only a small number of creators and sometimes generate more controversy than consensus. The reason is that one of the challenges associated with measuring the intrinsic quality of a creative work is how to formally define "quality."

In statistical modeling, this problem is typically addressed by positing the existence of latent (hidden) variables, which are unmeasurable but can be inferred from the values of other, measurable variables (1). For creative works, we presume there exists a latent variable, which we call "significance." Significance can be thought of as the lasting importance of a creative work. Significant works stand the test of time through novel ideas or breakthrough discoveries that change the landscape of a field or culture. Under this perspective, what is usually called "quality" is not the actual value of the latent variable, but an individual's or group's estimation of that value. Not surprisingly, the subjective evaluation of the unmeasurable true significance of the work is controversial, dependent on the historical moment, and very much "in the eye of the beholder."

Alternative methods for estimating the significance of a creative work fall under the labels of "impact" and "influence." Impact may be defined as the overall effect of a creative work on an individual, industry, or society at large, and it can be measured as sales, downloads, media mentions, or other possible means. However, in many cases, impact may be a poor proxy for significance. For example, *Duck Soup* (2) is generally considered to be the Marx Brothers' greatest film, but it was a financial disappointment for Paramount Pictures in 1933 (3). Influence may be defined as the extent to which a creative work is a source of inspiration for later works. Although this perspective provides a more nuanced estimation of significance, it is also more difficult to measure. For example, Ingmar Bergman's influence on later film directors is undebatable (4, 5), but not easily quantified. Despite different strengths and limitations, any quantitative approaches that result in an adequate estimation of significance should be strongly correlated when evaluated over a large corpus of creative works.

By definition, the latent variable for a creative work is inaccessible. However, for the medium of films—which will be the focus of this work—there is in fact as close to a measurement of the latent variable as one could hope for. In 1988, the US Government established the US National Film Preservation Board (NFPB) as part of the Library of Congress (6). The NFPB is tasked with selecting films deemed "culturally, historically, or aesthetically significant" for preservation in the National Film Registry (NFR). The NFR currently comprises 625 films "of enduring importance to American culture" (7). The careful evaluation and deliberation involved in the selection process each year, and the requirement of films being at least 10 y old to be eligible for induction, demonstrates the NFPB's true commitment to identifying films of significance.

Presence in the NFR is a binary variable as no distinctions are made between inducted films. This means that, although it can function as a "ground truth" for significances above a threshold value, it cannot discern the comparative significance of films. One of the goals of this study is to determine whether there are metrics that can accurately estimate film significance over a range of numerical values and for a large number of films. To this end, we investigate proxies of film quality, impact, or influence as potential measures of significance.

One can identify three main classes of approaches for estimating the significance of films: expert opinions, wisdom of the crowd, and automated methods. Expert opinions tend to measure the subjective quality of a film, whereas wisdom-of-the-crowd

Significance

Whether it is Hollywood movies or research papers, identifying works of great significance is imperative in a modern society overflowing with information. Through analysis of a network constructed from citations between films as referenced in the Internet Movie Database, we obtain several automated metrics for significance. We find that the best automated method can identify significant films, represented by selection to the US National Film Registry, at least as well as the aggregate rating of many experts and far better than the rating of a single expert. We hypothesize that these results may hold for other creative works.

Author contributions: M.W. and L.A.N.A. designed research; M.W. and X.H.T.Z. performed research; M.W., X.H.T.Z., and L.A.N.A. analyzed data; and M.W. and L.A.N.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1] To whom correspondence should be addressed. Email: amaral@northwestern.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1412198112/-/DCSupplemental.

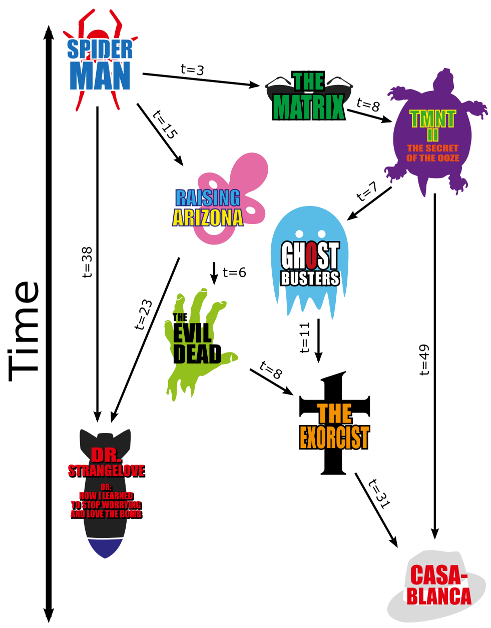

Fig. 1. Subgraph of film connections network. Films are ordered chronologically, based on year of release, from bottom to top (not to scale). A connection between two films exists if a sequence, sentence, character, or other part of the referenced film has been adopted, used, or imitated in the referencing film. For example, there is a connection from 1987's *Raising Arizona* (22) to 1981's *The Evil Dead* (23) because the main characters of both films drive an Oldsmobile Delta 88. Values represent the time lag of the connection, measured in years.

approaches tend to produce metrics that measure impact or popularity through broad-based surveys. Ideally, we can obtain an automated method that can measure influence. However, the best-known automated methods for films pertain to economic impact, such as the opening weekend or total box office gross. More recently, researchers and film industry professionals have evaluated films using electronic measures, such as Twitter mentions (8) and frequency of Wikipedia edits (9), but these may also be better indicators of impact or popularity. For an automated, objective measure that pertains to a film's influence, we turn to scientific works for an appropriate analog.

The network formed by citations from scientific literature is at the center of much research (10–12). Although some of the research on the scientific citation network aims to answer questions on citations between academic fields (13) or sex bias in academia (14), much work seeks to determine who is "winning" at science (15). Researchers have identified numerous metrics that are said to determine which paper (16), researcher (17), or journal (18) is the best, most significant, or most influential. These metrics range from the simple, such as total number of citations (19), to the complex, such as PageRank (20). The scientific citation network provides large quantities of data to analyze and dissect (12, 15, 21). If it were not for the expectation that researchers cite relevant literature, these metrics and indeed this avenue of study would not exist.

Like scientists, artists are often influenced or inspired by prior works. However, unlike researchers, artists are typically not obligated to cite the influences on their work. If data identifying citations between creative works could be made or obtained, we then could apply citation-based analyses to develop an objective metric for estimating the significance of a given work. As it happens, such data now exists. The Internet Movie Database (IMDb) (www.imdb.com) holds the largest digital collection of metadata on films, television programs, and other visual media. For each film listed in IMDb, there are multiple sections, from information about the cast and crew to critic reviews and notable quotes. Nestled among the deluge of metadata for each film is a section titled "connections," which contains a list of references and links to and from other films (Fig. 1). By analyzing this citation network obtained from user-edited data, we can investigate the suitability of metrics to estimate film significance based on the spread of influence in the world of motion pictures.

Data

In the network of film connections, a link from one film to another signifies that the former cites the latter in some form (24). For all citations in the network, the referencing film was released in a later calendar year than the film it references. Thus, the network contains no links that are "forward" or "sideways" in time. To account for sources of bias, we consider the giant component of the network of films produced in the United States (24). This subnetwork consists of 15,425 films connected by 42,794 citations.

We first compare the ratings obtained using various metrics from the three classes of estimation approaches (Table 1). For the expert opinions class, we have the choice of critic reviews

Table 1. Approaches for estimating the significance of films

Class	Method	Property	Strengths	Weaknesses
Expert opinions	Preservation board (e.g., NFR)	Significance	Consistent selection process Careful deliberation	Binary value Long time delay
	Critic reviews (e.g., Roger Ebert)	Quality	Subjective Many independent samples	Poor data availability Limited value range
	Awards (e.g., Oscars)	Quality	Distinctive Information for older items	Affected by promotion Restricted to small subset of films
Wisdom of the crowd	Average rating (e.g., IMDb user rating)	Quality/impact	Quantitative	Rater biases Unknown averaging procedure
	Total vote count (e.g., IMDb user votes)	Impact	Simple Quantitative	Proxy for popularity
Automated/objective measures	Economic measures (e.g., box office gross)	Impact	Quantitative	Proxy for popularity Data availability
	Electronic measures (e.g., Wikipedia edits)	Impact	Quantitative	Proxy for popularity Complex interpretation
	Citation measures (e.g., PageRank)	Influence	Quantitative	Complex interpretation

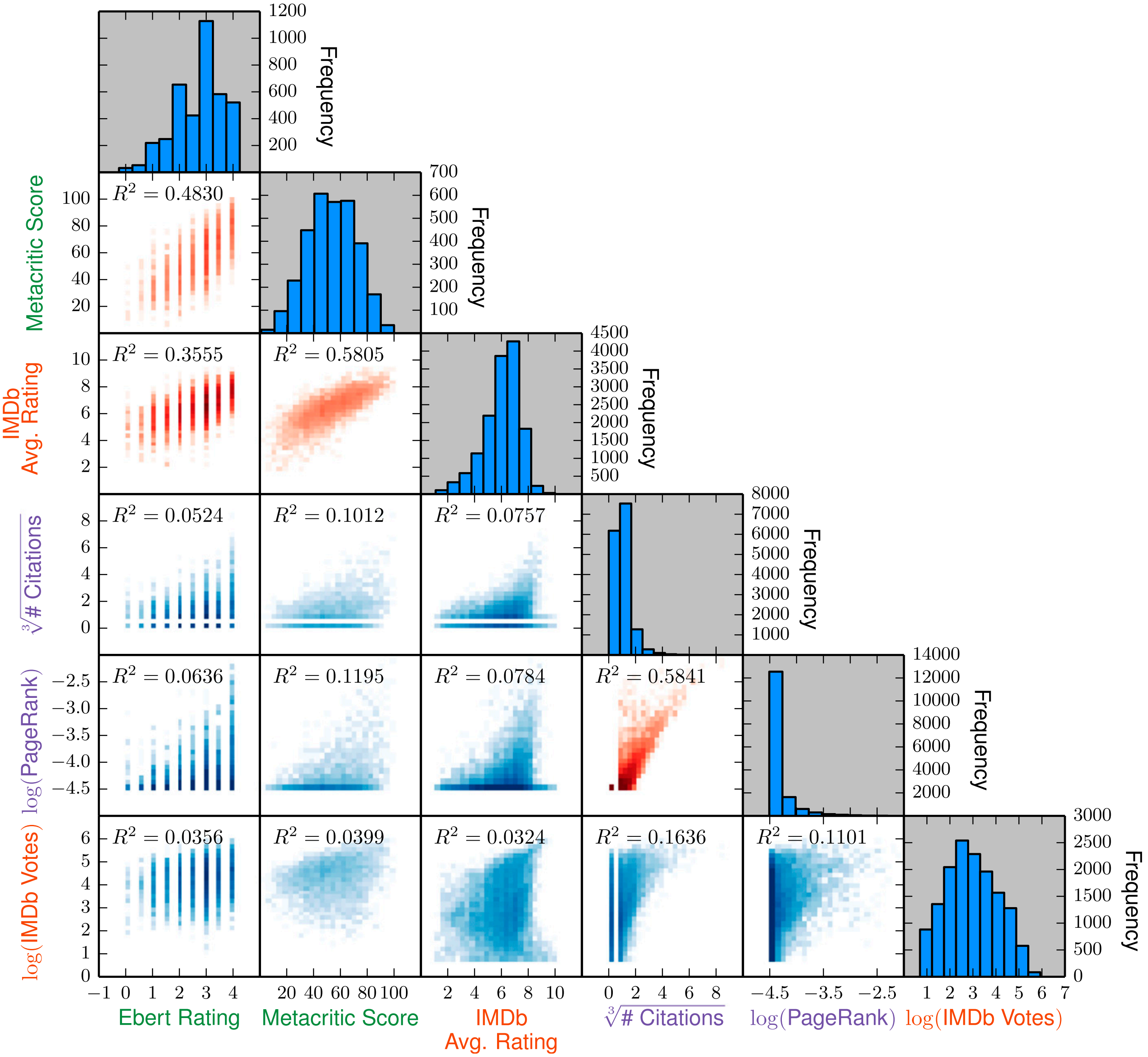

Fig. 2. Correlations and distributions of several estimators of significance. Plots with gray backgrounds are histograms. Plots with white backgrounds are scatter density plots depicting relationships between each pair of metrics (Roger Ebert star rating, Metacritic score, IMDb user rating, citation count, PageRank score, and total votes on IMDb). Adjusted R^2 values from linear regression analyses are shown for each pair of metrics. Stronger regressions ($R^2 > 0.25$) are depicted with a red gradient.

and artistic awards. For films, one of the strengths of critic reviews is that there are numerous independent samples. However, it is difficult to obtain reviews for older films by the same critic for all movies released since the beginnings of the industry (Fig. S1). Lack of data for older films is less of a concern for artistic awards, such as the Oscars, which date back to 1929. However, despite the great distinction of the Academy Awards, nominations are only given to a small subset of films, and wins to an even smaller subset. In addition, the Oscars are often affected by film popularity and studio promotion, which raises concerns about their accuracy in rewarding truly deserving films. For these reasons, we opt not to include award statistics in our analysis. Instead, we choose to consider two types of critic reviews: the star ratings of noted late film critic Roger Ebert and the aggregate critic review score reported by Metacritic. We include the former because of his long history as a renowned film critic. We include the latter because it provides a simple and self-consistent way to incorporate the ratings of multiple critics.

Population-wide surveys—a class that includes online polls—are well-suited for analysis as they are quantitative methods derived from large numbers of subjective opinions. This class of methods may be limited in identifying significance, however, due to biases and lack of expertise on the part of raters. The two population-wide survey metrics we analyze are the average IMDb user rating and the total number of user votes received on IMDb.

Finally, we consider two well-known statistics obtained from the connections network: total citations and PageRank score (25). Comparison of the six aforementioned statistics reveals that some of them exhibit moderate correlation (Fig. 2).

Table 2. Binary regression and Random Forest classification results for several estimators of significance

	Probit regression					Random Forest[†]	
Metric[*]	Fraction reported	Reported in NFR	Balanced accuracy[‡]	AUC[§]	$\mathrm{p}R^2$[¶]	Variable importance	Variable importance
Ebert rating[#]	0.242	0.061	0.5 (0.)	0.87 (0.01)	0.04 (0.01)	0.0070 (0.0019)	0.0043 (0.0012)
Metacritic score[#]	0.134	0.045	0.5 (0.)	**0.93** (0.01)	0.06 (0.02)	**0.0262** (0.0034)	**0.0235** (0.0034)
IMDb average rating[#]	0.957	0.039	0.502 (0.004)	**0.88** (0.01)	0.12 (0.01)	0.0217 (0.0051)	0.0186 (0.0042)
IMDb votes[#]	0.957	0.039	0.5 (0.01)	0.76 (0.01)	0.04 (0.01)	0.0103 (0.0017)	0.0078 (0.0012)
Total citations[‖]	1.000	0.037	0.57 (0.01)	0.86 (0.01)	**0.19** (0.02)	0.0201 (0.0031)	0.0133 (0.0018)
PageRank[‖]	1.000	0.037	**0.57** (0.01)	0.85 (0.01)	0.19 (0.02)	**0.0256** (0.0039)	0.0165 (0.0026)
Long-gap citations**	1.000	0.054	**0.61** (0.01)	0.88 (0.01)	**0.26** (0.02)	—	**0.0254** (0.0032)

[*] SDs in parentheses. Top two values for each performance category in bold.

[†] Cross-validated Random Forest classification performed on subset of 766 films with full information released on or before 1999.

[‡] Obtained from classification table analysis with 0.5 as the threshold.

[§] Area under the receiver operating characteristic (ROC) curve (29).

[¶] Tjur's pseudo-R^2 (30).

[#] Regression with Heckman correction performed on 12,339 films released on or before 2003. Used in both Random Forest analyses.

[‖] Regression performed on 12,339 films released on or before 2003. Used in both Random Forest analyses.

**Regression performed on 8,011 films released on or before 1986. Used only in second Random Forest analysis.

Results

We conduct a probit regression analysis of the dependent binary variable indicating whether or not a film is in the NFR, using the Heckman correction method (26, 27) to account for missing data. We also perform Random Forest classification (28) using the six metrics as predictors and selection to the NFR as the response (Table 2). To avoid overfitting, our Random Forest analysis is cross-validated by running 100 trials with 80% of the data points chosen at random without replacement. In addition, we use a multivariate probit regression model incorporating all of the metrics discussed so far (Table 3).

We find that Metacritic score is a far more important variable than the Roger Ebert rating at indicating presence in the NFR based on Random Forest classification. The Metacritic score probit model also outperforms the Ebert rating model in terms of area under the curve. Thus, single-expert ratings do not appear to identify significant films as well as an aggregation of expert ratings. Also, the automated metrics—total citation count and PageRank—perform much better than single-expert evaluation and at least as well as IMDb average ratings. Between the two, PageRank is more important in Random Forest classification (Table 2), whereas total citation count is a better fit in the multivariate probit model, where it accounts for more of the correlation than all other variables (Table 3).

Note that these results must be interpreted with some caution. In particular, Metacritic score is predisposed to perform better in analyses in which we do not account for missing data, such as Random Forest classification. This is due to significantly fewer data points in the subset of films considered, as fewer than 15% of films released before 1995 have a Metacritic score (Fig. S1). The few films from that period with Metacritic scores are more likely to have been rereleased and to be classics, and thus have high ratings from reviewers. This fact is made quantitative by the low balanced accuracy for the Metacritic score model when applying the Heckman correction (Table 2). Ignoring missing data in performing the probit regression yields a much higher (but misleading) balanced accuracy for both Metacritic score and Ebert rating (Table S1).

Although the automated methods perform well, we hypothesize that their performance could be further improved. Indeed, it is plausible that not all citations are the same. Thus, we next investigate the distribution of the "time lag" of edges in the connections network. The time lag of an edge is the number of years between the release of the edge's citing film and the release of the edge's cited film (Fig. 1). As an example, the edge linking *When Harry Met Sally…* (1989) (32) to *Casablanca* (1942) (33) has a time lag of 47. Note that given our rules for constructing the network, all time lag values are strictly positive.

Naïvely, one would expect that the frequency of connections as a function of time lag decreases monotonically, as new films would likely reference films released shortly before due to those films' shared cultural moment. Indeed, connections with a time lag of 1 y are the most numerous in the network, and for the most part, frequency of connections does decrease as time lag increases (Fig. 3 *A* and *B*). However, the distribution shows a surprising uptick for time lags around 25 y.

To explain this nonmonotonicity, we compare the data to two null models. The first null model is the "base" or "unbiased" null model wherein connections in the network are randomly redirected (34, 35). The second is a "biased" null model wherein connections are randomly redirected, but with a preference toward creating connections with shorter time lags. For both null models, we assume that all films retain the same number of links in and out, and, as with the actual film connections network, there are no back-in-time citations (Fig. S2).

We find that the unbiased null model mimics the time lag distribution for values greater than 22 y, but it fails to predict the distribution for values less than 22 y (Fig. 3*A*). In contrast, the biased null model accurately predicts the time lag distribution for values between 2 and 18 y, but is not consistent with the data for time lags greater than 19 y (Fig. 3*B*).

The citation trend of recent films, wherein they are cited more often than expected by an unbiased null model, is not a result of the sizable growth of films in IMDb in the past several years. We

Table 3. Contributions of several estimators of significance in multivariate probit regression (see also Table S2)

Model	$\mathrm{p}R^2$*	$\Delta\mathrm{p}R^2$
Metacritic + IMDb rating +		
IMDb votes + total citations	0.6063	—
– Total citations	0.4856	−0.1207
– IMDb votes	0.5411	−0.0652
– Metacritic	0.5432	−0.0631
– IMDb rating	0.5548	−0.0515
Metacritic + IMDb rating +		
Long-gap citations	0.6246	—
– Long-gap citations	0.4805	−0.1441
– Metacritic	0.5572	−0.0674
– IMDb rating	0.5848	−0.0398

*McFadden's pseudo-R^2 (31).

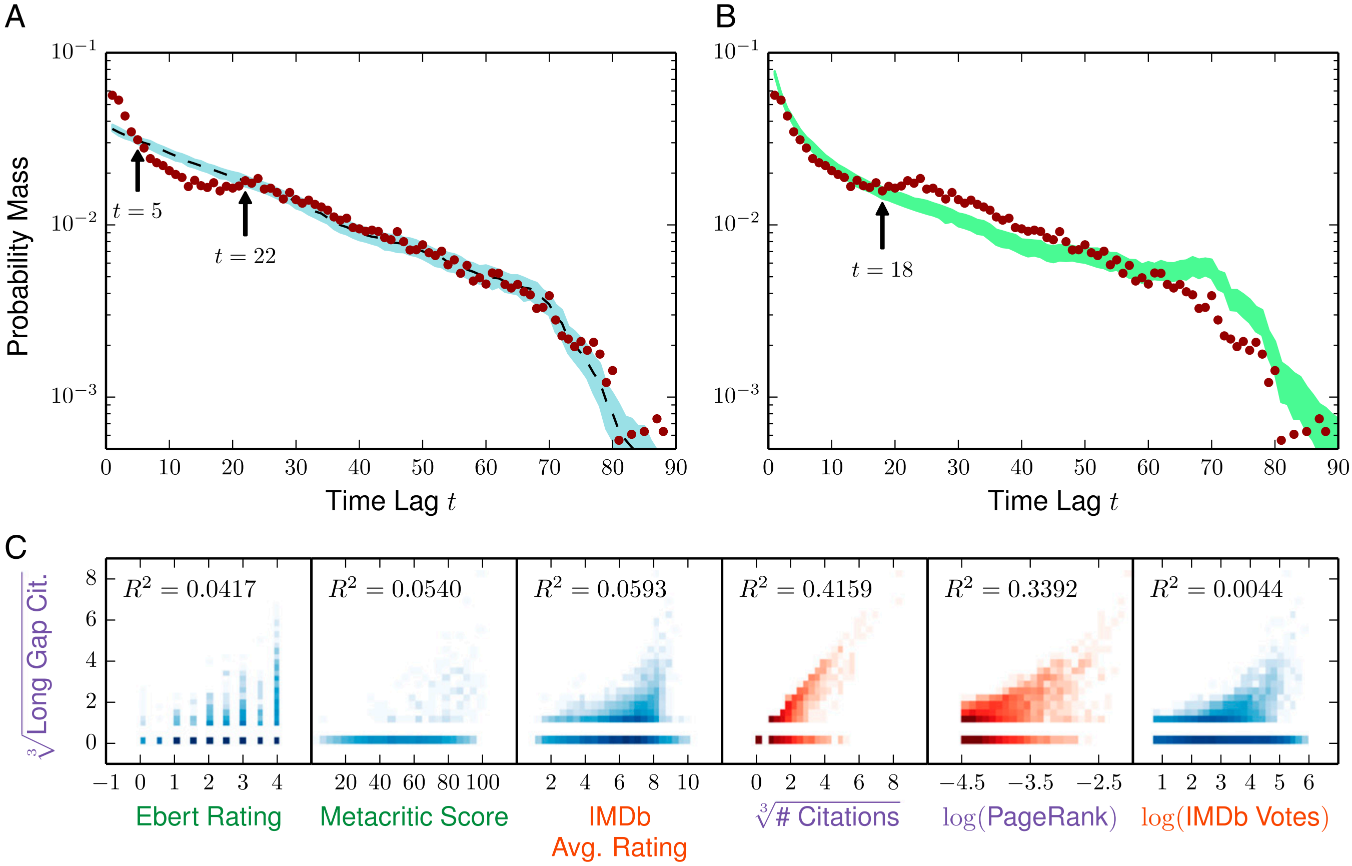

Fig. 3. Null distributions of time lag and correlations involving long-gap citations. (*A* and *B*) Shaded regions are 95% confidence intervals for the null models resulting from random rewiring of the network. The shaded blue region (*A*) is for the unbiased null model. The shaded green region (*B*) is for the null model with a bias toward links with shorter time lags. The dashed black line (*A*) is the theoretical average distribution of the unbiased null model (*SI Text*, Eq. **S3**). Arrows identify the values where the actual distribution diverges from the null models. (*C*) Scatter density plots depicting relationships between long-gap citation count and the other metrics. Adjusted R^2 values are shown. Stronger regressions are depicted with a red gradient.

find that this result persists even if we omit all films made after 2000, after 1990, and after 1970 (Fig. S3).

The accuracy of the biased null model for shorter time lags indicates the likelihood of many films receiving shorter-gap citations (fewer than 20 y). However, the frequency of these citations quickly falls off with time for most films that receive them. The accuracy of the unbiased null model for longer time lags suggests that, for certain films, timeliness does not matter. We presume that films that receive these long time lag citations (25 y or more) may be considered more significant as they continue to be cited regardless of time.

Prompted by these modeling results, we investigate the possibility that one can use the total count of "long-gap citations," our term for citations received with a time lag of 25 y or more, as a proxy for significance. To determine whether long-gap citation count is an accurate estimator in this regard, we compare its performance to that of the other metrics we have previously considered. We find that long-gap citation count correlates reasonably well with PageRank and total citation count, but not with the nonautomated metrics (Fig. 3*C*).

Our analysis shows that long-gap citation count is a strong predictor for presence in the NFR (Tables 2 and 3). Random Forest analysis yields that long-gap citation count is the most important predictor of NFR presence when incorporated with all other metrics, ahead of Metacritic score. Importantly, the long-gap citations model consistently outperforms both PageRank and total citations. This indicates that long-gap citation count is a superior identifier of significant films compared with other metrics.

An aspect of all of the analyses performed so far is that one cannot differentiate between highly rated films that are significant in their entirety versus films that are significant because of an iconic moment. Fortunately, many of the connections listed on IMDb include a brief note describing the specific link between the films. For a limited set of films—the 15 films with long-gap citation counts between 51 and 60 (Table S3)—we manually classify their citations by description and determine to what extent the citation covers each film, either broadly or for just a single aspect (Table S4). We thereby see that 55% of annotated citations of *The Seven Year Itch* (36) reference the famous scene where Marilyn Monroe's white dress blows up from the passing subway and that 35% of annotated citations of *North by Northwest* (37) reference the crop duster scene. We also observe that 71% of annotated citations of *Bride of Frankenstein* (38) and 70% of annotated citations of *Mary Poppins* (39) reference the entire film or the title character. Our analysis of these 15 films suggests that some films are indeed significant because of iconic scenes or characters.

To extend this type of analysis to the entire set of films, we consider a number of metrics that reflect the similarity present in the citation descriptions for a film. Unfortunately, we find no correlation with the aforementioned percentage values (Fig. S4) and are thus unable to draw broad conclusions on this matter. It is certainly possible that many of the filmmakers citing *The Seven Year Itch* or *Bride of Frankenstein* have never actually seen the film they are referencing, but that underlines how much the famous dress and the memorable hair are firmly engrained in popular culture, that is, how significant at least these moments in these movies truly are.

Discussion

Our cross-evaluation of metrics to estimate the significance of movies uncovers two main findings. First, aggregation of numerous expert ratings performs better as an indicator of significant films than the ratings of an individual expert. Our second and more important result is that well-conceived automated methods can perform as well as or better than aggregation of expert opinions at identifying significant films, even when we do not account for missing rating data. Not only are automated methods superior identifiers of significance, they are the most scalable for application to large numbers of works.

Although our work pertains to films, it is not unconceivable that these same insights may hold for other creative enterprises, including scientific research. It is well within the realm of possibility that a well-designed automated method, potentially rooted in network analysis, can outperform even the best experts at identifying the most significant scientific papers.

Our examination of the network of IMDb film connections reveals additional insights about how ideas and culture spread over time. There is a clear preference for current films to make references to films from the recent past. Although this seems intuitive, the fact that films released within the prior 3 y are referenced at a higher rate than expected from an unbiased null model is surprising. It suggests that the film industry relies heavily on recently popular ideas when making new films. It is also possible that this trend reflects the public's focus on what is "new and fresh."

Because the distribution of time lag begins aligning with the unbiased null model at 25 y, it implies that the significant films from any given year will be definitively known once 25 y have passed, as those films will be the ones that continue to receive citations. This is verified by the strong correlation between the long-gap citation count of a film and its presence in the NFR. However, long-gap citation counts not only identify instantly notable films such as *Star Wars* (40) and *Casablanca*, but also films that were not immediately appreciated. For example, *Willy Wonka & the Chocolate Factory* (41) was a box office disappointment when it was released in 1971 (42). However, the film gained a significant following a decade later thanks to home video sales and repeated airings on cable television and is today considered a top cult classic (42). The story behind *Willy Wonka* is reflected in the film connections network: it has no citations with a time lag of 4 y or less, but 52 long-gap citations, the 37th-most of all films in our analysis (Table 3). Interestingly, *Willy Wonka* is not currently in the NFR, but that does not mean it will not be added at a later date. *Mary Poppins*, which has the 33rd-most long-gap citations, was only added in 2013, nearly 50 y after its release (7). Likewise, *Dirty Harry* (43)—released the same year as *Willy Wonka* and having accrued 51 long-gap citations—was not inducted until 2012.

Twenty-five years may seem like a long time to wait before we can begin quantifying film significance. However, significance by definition may not be readily apparent. This is true of other forms of art, as well as any other field where influence spreads. There is a reason the Nobel Prize is no longer awarded for research done in the same year (44). A film's significance should ultimately be judged on how its ideas influence filmmaking and culture in the long term.

ACKNOWLEDGMENTS. We thank Adam Hockenberry; Alan Wasserman, MD; Andrea Lancichinetti; Arnau Gavalda, Emöke-Ágnes Horvát; Filippo Radicchi; Irmak Sirer; João Moreira; Julia Poncela; and Konrad Körding for their comments, suggestions, and insights. We also thank the reviewers and editor for helpful suggestions and comments. We thank Penny Dash for editorial assistance. This work was partially supported by Army Research Office Grant W911NF-14-1-0259 (to L.A.N.A.).

1. Borsboom D, Mellenbergh GJ, van Heerden J (2003) The theoretical status of latent variables. *Psychol Rev* 110(2):203–219.
2. McCarey L (Director) (1933) *Duck Soup* [Motion picture] (Paramount Pictures, Hollywood, CA).
3. Louvish S (1999) *Monkey Business: The Lives and Legends of the Marx Brothers* (Faber and Faber, London).
4. Corliss R (2007) Woman, man, death, god. *Time* 170(7):65.
5. Macnab G (2009) *Ingmar Bergman: The Life and Films of the Last Great European Director* (I. B. Tauris, London).
6. Library of Congress (2014) National Film Preservation Board. Available at www.loc.gov/film/index.html. Accessed April 11, 2014.
7. Library of Congress (2014) National Film Registry. Available at www.loc.gov/film/filmnfr.html. Accessed April 11, 2014.
8. Rui H, Liu Y, Whinston A (2013) Whose and what chatter matters? *Decis Support Syst* 55(4):863–870.
9. Mestyán M, Yasseri T, Kertész J (2013) Early prediction of movie box office success based on Wikipedia activity big data. *PLoS One* 8(8):e71226.
10. Price DJ (1965) Networks of scientific papers. *Science* 149(3683):510–515.
11. de Solla Price DJ (1976) A general theory of bibliometric and other cumulative advantage processes. *J Am Soc Inf Sci* 27(5):292–306.
12. Newman MEJ (2003) The structure and function of complex networks. *SIAM Rev Soc Ind Appl Math* 45(2):167–256.
13. Chen C, Hicks D (2004) Tracing knowledge diffusion. *Scientometrics* 59(2):199–211.
14. Duch J, et al. (2012) The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PLoS One* 7(12):e51332.
15. Redner S (1998) How popular is your paper? An empirical study of the citation distribution. *Eur Phys J B* 4(2):131–134.
16. Redner S (2005) Citation statistics from 110 years of *Physical Review*. *Phys Today* 58(6):49–54.
17. Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: Toward an objective measure of scientific impact. *Proc Natl Acad Sci USA* 105(45):17268–17272.
18. Garfield E (2006) The history and meaning of the journal impact factor. *JAMA* 295(1):90–93.
19. Garfield E (1972) Citation analysis as a tool in journal evaluation. *Science* 178(4060):471–479.
20. Chen P, Xie H, Maslov S, Redner S (2007) Finding scientific gems with Google's PageRank algorithm. *J Informetrics* 1(1):8–15.
21. Seglen PO (1992) The skewness of science. *J Am Soc Inf Sci* 43(9):628–638.
22. Coen E (Producer), Coen J (Director) (1987) *Raising Arizona* [Motion picture] (20th Century Fox, Los Angeles).
23. Tapert R (Producer), Raimi S (Director) (1981) *The Evil Dead* [Motion picture] (New Line Cinema, Los Angeles).
24. 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.
25. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30(1-7):107–117.
26. 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.
27. Heckman JJ (1979) Sample selection bias as a specification error. *Econometrica* 47(1):153–161.
28. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32.
29. Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clin Chem* 39(4):561–577.
30. Tjur T (2009) Coefficients of determination in logistic regression models—a new proposal: The coefficient of discrimination. *Am Stat* 63(4):366–372.
31. McFadden D (1974) *Conditional Logit Analysis of Qualitative Choice Behavior. Frontiers in Econometrics*, ed Zarembka P (Academic, New York), pp 105–142.
32. Ephron N (Producer), Reiner R (Producer and Director), Scheinman A (Producer) (1989) *When Harry Met Sally*... [Motion picture] (Columbia Pictures, Culver City, CA).
33. Wallis HB (Producer), Curtiz M (Director) (1942) Casablanca [Motion picture] (Warner Bros., Burbank, CA).
34. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. *Science* 296(5569):910–913.
35. 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.
36. Feldman CK (Producer), Wilder B (Producer and Director) (1955) *The Seven Year Itch* [Motion picture] (20th Century Fox, Los Angeles).
37. Hitchcock A (Director) (1959) *North by Northwest* [Motion picture] (Metro-Goldwyn-Mayer, Beverly Hills, CA).
38. Laemmle C, Jr (Producer), Whale J (Director) (1935) *Bride of Frankenstein* [Motion picture] (Universal Pictures, Universal City, CA).
39. Disney W (Producer), Stevenson R (Director) (1964) *Mary Poppins* [Motion picture] (Buena Vista Distribution, Burbank, CA).
40. Kurtz G (Producer), Lucas G (Director) (1977) Star Wars [Motion picture] (20th Century Fox, Los Angeles).
41. Margulies S (Producer), Wolper DL (Producer), Stuart M (Director) (1971) *Willy Wonka & the Chocolate Factory* [Motion picture] (Warner Bros., Burbank, CA).
42. Stuart M (2002) *Pure Imagination: The Making of Willy Wonka and the Chocolate Factory* (St. Martin's, New York).
43. Siegel D (Producer and Director) (1971) Dirty Harry [Motion picture] (Warner Bros., Burbank, CA).
44. Pettersson R (2001) The Nobel Prizes in the new century. An interview with Ralf Pettersson, Director of the Stockholm Branch of the Ludwig Institute for Cancer Research, the Karolinska Institute, and former chairman of the Nobel Prize Committee for Physiology/Medicine. Interview by Holger Breithaupt. *EMBO Rep* 2(2):83–85.

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:

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

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:

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

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:

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

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:

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

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:

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

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:

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

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.

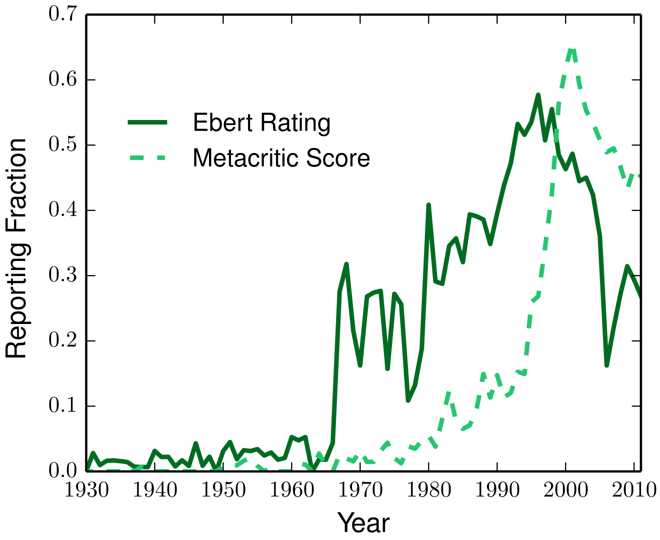

Fig. S1. Fraction of reported critic data values by year.

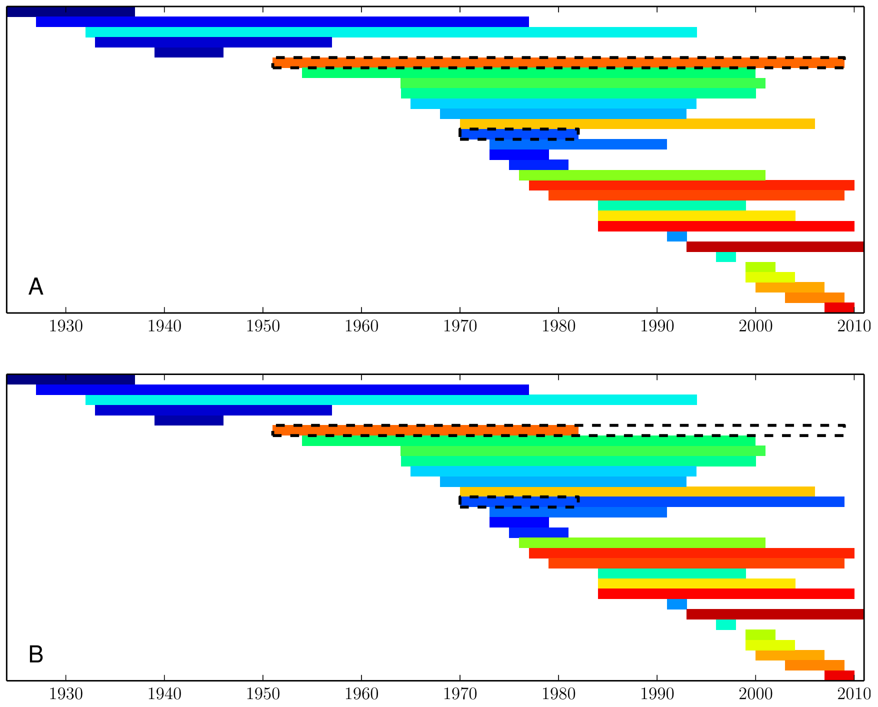

Fig. S2. One step of rewiring simulation. The diagram depicts a randomly chosen sample of edges from the directed network of film connections. Each bar represents an edge in the sample. The right end of each bar indicates the year of release for the citing film. The left end of each bar indicates the 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.

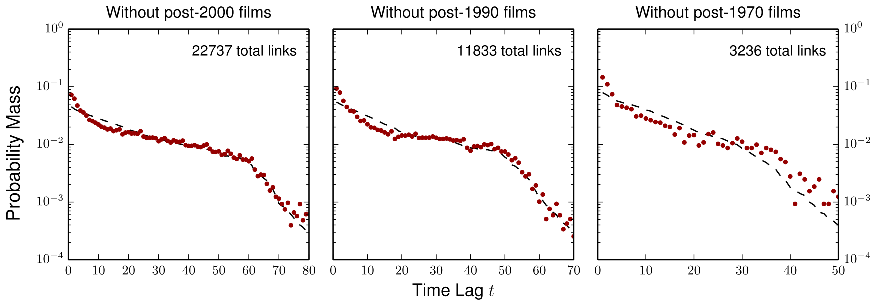

Fig. S3. Distribution of time lag with exclusions. Probability mass function of the time lag of connections in the film connections network, discounting all films made after 2000, after 1990, and after 1970. Brown points represent the actual distributions. Dashed black lines represent the unbiased null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

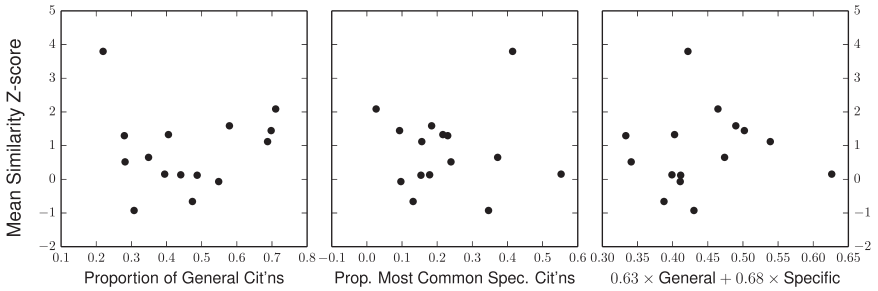

Fig. S4. Mean similarity *Z* score versus annotated citation classifications. Scatterplots comparing the mean similarity *Z* scores of citation annotations to the fractions of annotations under certain classifications for 15 highly cited films (See Table S4). The fractions considered are the proportion 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.