

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 ^aEngineering Sciences and Applied Mathematics, ^bChemical and Biological Engineering, and ^cPhysics and Astronomy, ^dHoward Hughes Medical Institute, and ^eNorthwestern 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.

¹To 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.

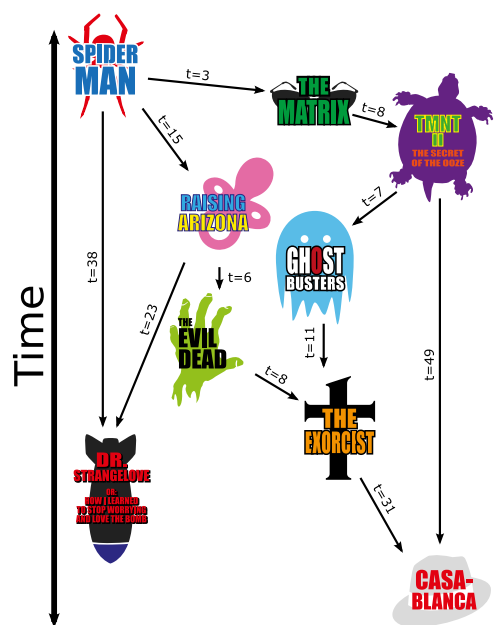


Fig. 1. Subgraph of film connections network. Films are ordered chronologically, based on year of release, from bottom to top (not to scale). A connection between two films exists if a sequence, sentence, character, or other part of the referenced film has been adopted, used, or imitated in the referencing 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.

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

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

Metric*	Probit regression					Random Forest [†]	
	Fraction reported	Reported in NFR	Balanced accuracy [‡]	AUC [§]	pR ^{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-R² (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. 3A). 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. 3B).

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	pR ^{2*}	ΔpR 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-R² (31).

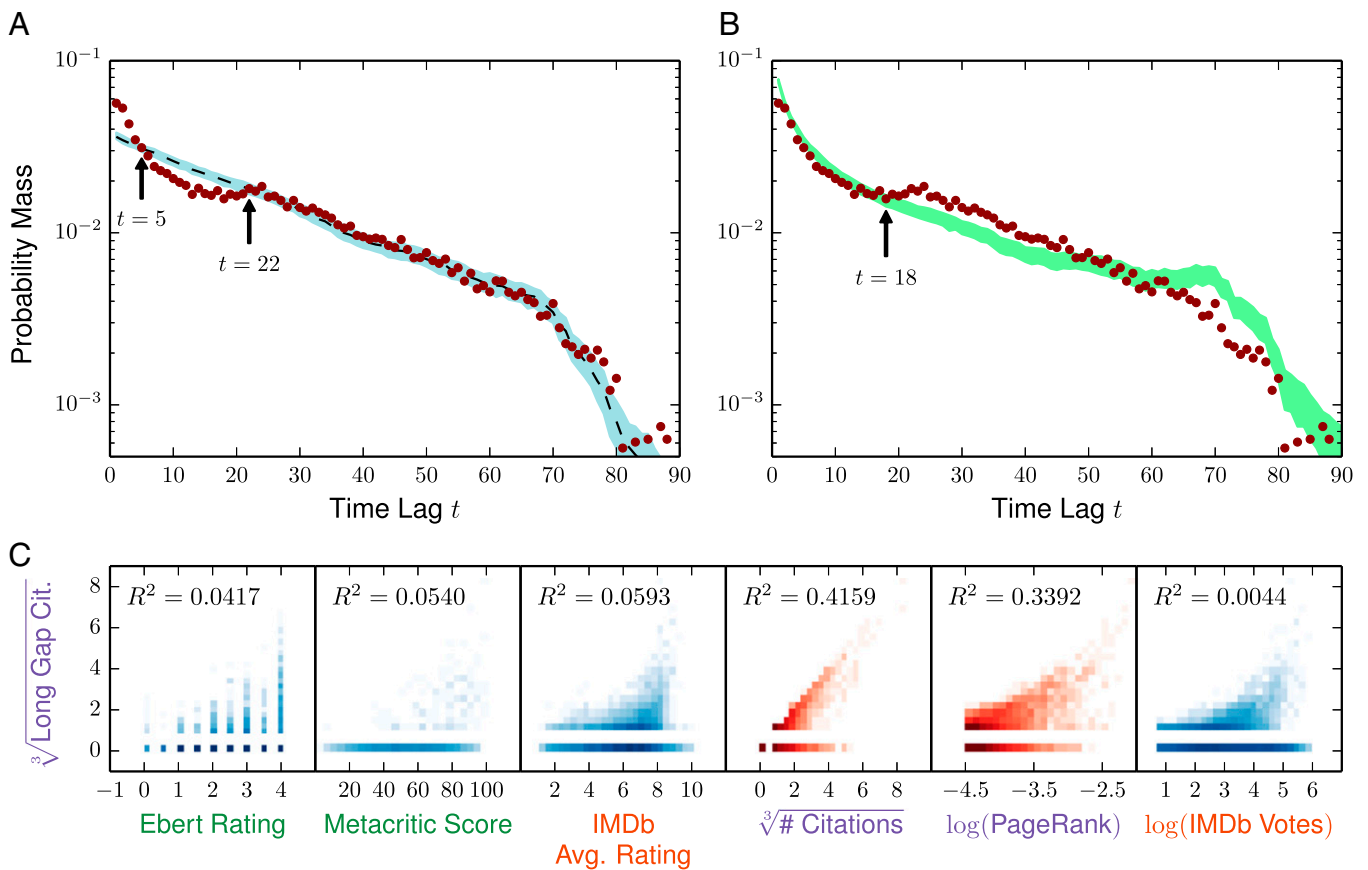


Fig. 3. Null distributions of time lag and correlations involving long-gap citations. (A and B) Shaded regions are 95% confidence intervals for the null models resulting from random rewiring of the network. The shaded blue region (A) is for the unbiased null model. The shaded green region (B) is for 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. 3C).

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 Gavaldà, Emöke-Ágnes Horvát; Filippo Radicchi; Irmak Sırer; 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. Petterson R (2001) The Nobel Prizes in the new century. An interview with Ralf Petterson, 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.