

RESEARCH ARTICLE

Statistical Patterns in Movie Rating Behavior

Marlon Ramos¹, Angelo M. Calvão¹, Celia Anteneodo^{1,2}*

1 Department of Physics, PUC-Rio, Rio de Janeiro, Brazil, **2** National Institute of Science and Technology for Complex Systems, Rio de Janeiro, Brazil

* These authors contributed equally to this work.

* celia.fis@puc-rio.br

Abstract

Currently, users and consumers can review and rate products through online services, which provide huge databases that can be used to explore people's preferences and unveil behavioral patterns. In this work, we investigate patterns in movie ratings, considering IMDb (the Internet Movie Database), a highly visited site worldwide, as a source. We find that the distribution of votes presents scale-free behavior over several orders of magnitude, with an exponent very close to $3/2$, with exponential cutoff. It is remarkable that this pattern emerges independently of movie attributes such as average rating, age and genre, with the exception of a few genres and of high-budget films. These results point to a very general underlying mechanism for the propagation of adoption across potential audiences that is independent of the intrinsic features of a movie and that can be understood through a simple spreading model with mean-field avalanche dynamics.



OPEN ACCESS

Citation: Ramos M, Calvão AM, Anteneodo C (2015) Statistical Patterns in Movie Rating Behavior. PLoS ONE 10(8): e0136083. doi:10.1371/journal.pone.0136083

Editor: Wei-Xing Zhou, East China University of Science and Technology, CHINA

Received: May 6, 2015

Accepted: July 29, 2015

Published: August 31, 2015

Copyright: © 2015 Ramos et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors acknowledge financial support from Brazilian Agencies CNPq, CAPES and Faperj. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

In recent decades, statistical physics has contributed to the study of social dynamics through theoretical models, providing insights and uncovering the crucial laws that govern many phenomena, such as the spread of information, rumors and opinions [1, 2]. While there has been notable progress in developing theoretical models, their validation by direct confrontation with real data has yet to be achieved. Thanks to the popularization of online social networks and, more recently, of websites for ratings and recommendations, new possibilities have arisen to explore this field. On one hand, ratings provide relevant information on how people's preferences are distributed. On the other hand, beyond practical applications aiming to improve recommender systems, empirical data allow checking or validating theoretical models that can then be further used to interpret and predict observed outcomes [3].

Here, we aim to explore patterns in movie rating behavior as a source of information on the distribution of people's preferences. We believe that analyzing the number of votes (where a vote consists of assigning a star rating) rather than, for example, the total number of movie admissions, is a suitable way to measure the popularity of a given movie.

People can watch a movie in many different ways; therefore, the number of admissions provides only partial information. Analyzing a movie's total gross also suffers from the same problem, and it is difficult to sum all the sources of movie income. Instead, the number of votes is

independent of the means used to watch a movie. Furthermore, it is a direct and accessible source of information as well as a much richer one, as we can also extract the public’s opinion about the film. Moreover, while some theoretical models indicate that the distribution of adoption cascades follow a power law [4–7], there is a lack of empirical evidence to validate these results. In this sense, the present work can contribute significantly.

Empirical results

We analyzed data from the Internet Movie Database (IMDb), a source of information about movies and related content that allows visitors to review and rate movies and other entertainment items online (S1 Dataset). It is one of the most visited sites worldwide and the first in its field [8]. We collected the number of votes received by each movie, where a vote consists of assigning a rating on a scale from 1 to 10 stars where 1 star means *awful* and 10 stars means *excellent*. We estimated the probability distribution $P(n_v)$ of the number of votes n_v by building the normalized histogram shown in Fig 1. A remarkable scale-free behavior over approximately four orders of magnitude of n_v emerges, with a power law exponent 1.51 ± 0.02 and an exponential cutoff. As a test for possible biases, we considered an alternative database provided by Netflix, and we observed the same scaling behavior (In October 2006 Netflix launched a competition to enhance the performance of its recommender system [9]. As part of the competition, they released a dataset that consists of approximately 100 million ratings by users on 17 thousand movies).

There are several processes that can generate power laws [10, 11], in particular, a statistical mixture of different scales. The use of a mixture of processes to explain the scaling behavior is, in fact, a possibility in this case, as the data include different categories of movies, e.g., TV movies, feature films, and items of different ages and with different ratings. Therefore, we should conduct a more refined analysis after separating the subsets by different criteria.

Rating

We begin by inspecting the relation between the number of votes n_i and the average rating $\langle r_i \rangle$ of each movie i . In Fig 2, we depict a scatter plot of n_i vs. $\langle r_i \rangle$. Although the points are very spread out, trends can be identified. The logarithmic scale on the ordinate axis implies a broad

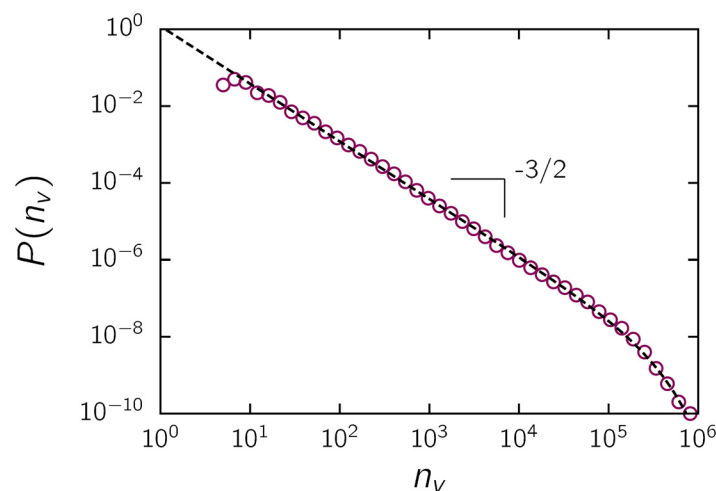


Fig 1. Distribution of the number of votes for IMDb movies (S1 Dataset). The dashed line fits the data of the function $P(n_v) = An_v^{-\alpha} \exp(-\lambda n_v)$, where A is a normalization constant, $\alpha = 1.51$ and $\lambda = 4.0 \cdot 10^{-6}$.

doi:10.1371/journal.pone.0136083.g001

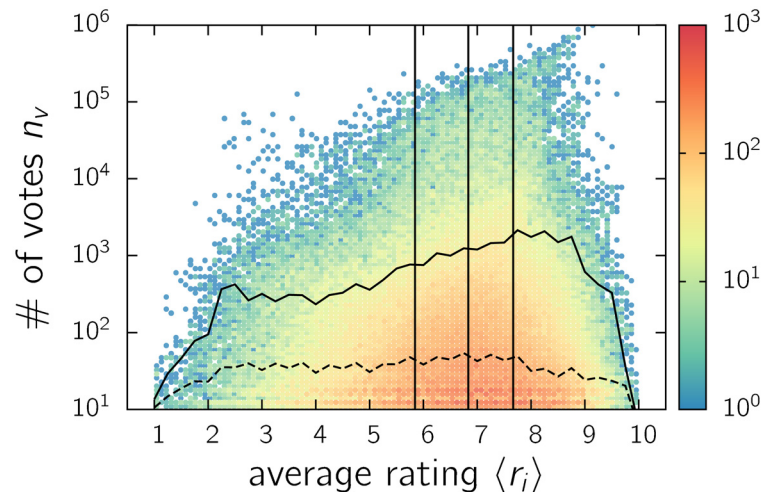


Fig 2. Color map of the number of votes n_i vs. the average rating $\langle r_i \rangle$ of each IMDb movie i . In the color map, each bullet contains the number of movies indicated by the color scale. The white region indicates zero movies. The dotted and dashed lines represent the (binned) arithmetic and geometric mean values, respectively. The vertical lines indicate the quartiles, which divide the dataset into four groups ($G_{4,1}, \dots, G_{4,4}$) of equal size.

doi:10.1371/journal.pone.0136083.g002

distribution in the number of votes for each bin of average ratings, as observed for the entire set (Fig 1). The distribution of votes within the range of average ratings is asymmetric, with a bias towards positive values, which differs from the expectations for a random profile. The maximal number of votes increases with the rating, indicating that an extremely large number of votes is associated with well-rated movies. A very similar picture has been reported for ratings on Yahoo music [12, 13].

The number of votes must increase with the number of people that watched a movie, which in turn is expected to be higher when more people like the movie. Moreover, more people liking a movie will drive higher ratings. Therefore, we would expect a positive correlation between ratings and the number of votes. Although the cutoff increases with the average rating, a high number of high-rated movies receive few votes. Moreover, the geometric mean of the number of votes as a function of the average rating presents a flat profile (Fig 2), although the arithmetic mean slightly increases because it is influenced by extreme values.

A natural question arises: how is the normalized distribution of votes affected by movie ratings? To examine this issue, we obtained $P(n_v)$ for separate groups of data, splitting the entire set by the median with respect to the rating. That is, we separately considered the lower- and higher-rated halves of the entire dataset, $G_{2,1}^r$ and $G_{2,2}^r$, respectively. We also considered the quartiles and subdivided the dataset into the groups $G_{4,1}^r, \dots, G_{4,4}^r$. The results are shown in Fig 3. Note that $P(n_v)$ is almost insensitive to whether a rating is favorable or not. That is, data above and below the median present the same pattern, coinciding over the four decades of the power law regime (Fig 3a). The main discrepancy appears at the exponential cutoff above 10^5 votes, where the decay occurs faster for the lower-rated half. This is consistent with the fact that low-rated movies do not receive the extremely high number of votes given to high-rated movies. However, the same scale-free behavior holds for both halves over several orders of magnitude in the number of votes, pointing to a mechanism independent of the attributes measured by ratings. The same tendencies are observed at the level of quartiles, as depicted in Fig 3b, and at the level of the number of stars (not shown).

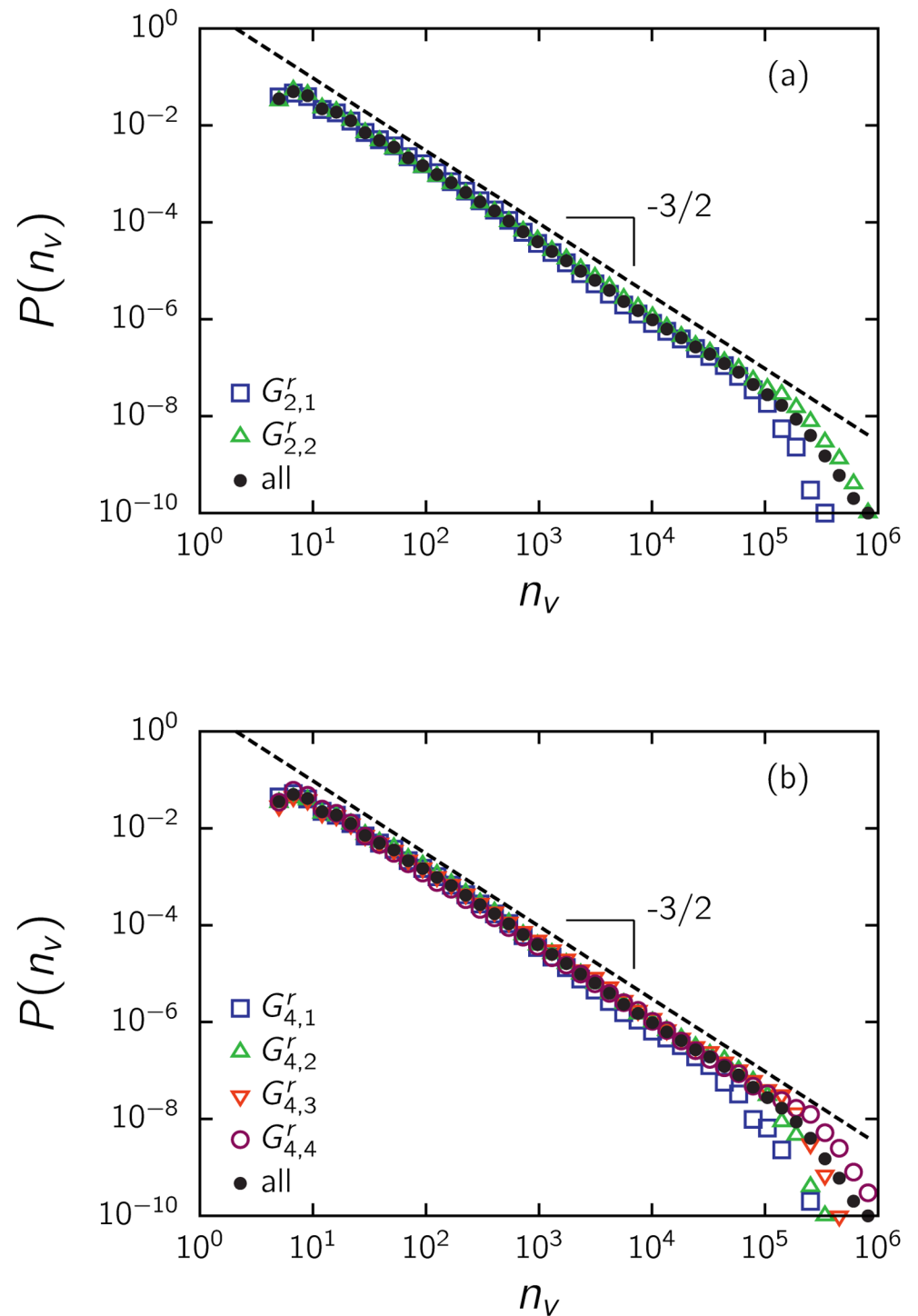


Fig 3. Impact of ratings on the distribution $P(n_v)$ of votes for IMDb movies, for (a) the two groups $G_{2,1}^r$ and $G_{2,2}^r$ separated by the median and (b) the four groups $G_{4,1}^r, \dots, G_{4,4}^r$ determined by the quartiles, as indicated in Fig 2. In this and other equivalent figures, the dashed line with slope $-3/2$ is drawn for comparison with the distribution of the entire dataset.

doi:10.1371/journal.pone.0136083.g003

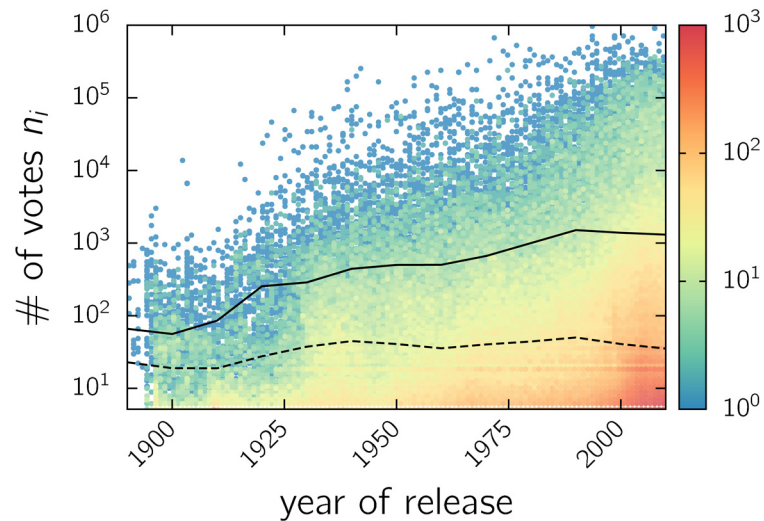


Fig 4. Color map of number of votes n_i vs. the year of release of IMDb movies. Each bullet contains the number of movies indicated by the color scale. The dotted and dashed lines represent the (binned) arithmetic and geometric mean values, respectively.

doi:10.1371/journal.pone.0136083.g004

Year of release

Second, we investigate the impact of the age of the items on $P(n_v)$. Fig 4 shows a scatter plot for the number of votes vs. the year of release for each movie. We note that the number of votes depends on the movie's age. The cutoff for the number of votes increases as age decreases. Concomitantly, younger movies tend to receive more votes on average. All these tendencies are consistent with the expectation that new movies receive substantially more votes than older ones. Furthermore, old movies can receive votes only retroactively (the IMDb user registration system was launched in 1997 [14]), while a new movie can receive votes contemporaneously with its more effervescent phase.

Again in this case, we analyzed $P(n_v)$ for subsets separated by the year of release. Sets of movies younger than y years, even those released within one year and hence with worse statistics, present the same pattern over the entire interval (see Fig 5a). To take an even closer look, we examined movies grouped by release time interval (Fig 5b). A significant discrepancy exists only for older movies, at the tail above 10^4 votes, in accordance with Fig 4, which shows that the cutoff occurs at a smaller n_v as age increases. However, the scaling region still spans more than three orders of magnitude of n_v , even in this case.

Genre

We also show $P(n_v)$ for TV series and feature movies separately (Fig 6). Both categories behave similar to the entire dataset, with a discrepancy occurring only at the cutoff.

We further divided the list of items by genre, considering only those genres with more than 15 thousand films (S1 Dataset). The two most numerous genres, comedy ($\alpha = 1.50 \pm 0.03$) and drama ($\alpha = 1.51 \pm 0.03$), have the same scaling as the entire dataset within error bars and practically coinciding over the entire range (Fig 7a). In addition, the same scaling holds within error bars (± 0.04 for the smaller sets) for animation ($\alpha \simeq 1.56$), family ($\alpha \simeq 1.55$), horror ($\alpha \simeq 1.45$) and romance ($\alpha \simeq 1.46$), as shown in Fig 7b. A slightly smaller exponent emerges for action ($\alpha \simeq 1.42$), adventure ($\alpha \simeq 1.38$), crime ($\alpha \simeq 1.40$) and thrillers ($\alpha \simeq 1.40$), with each category having fewer than 20 thousand films (S1 Dataset). Shorts ($\alpha \simeq 2.1$) and documentary

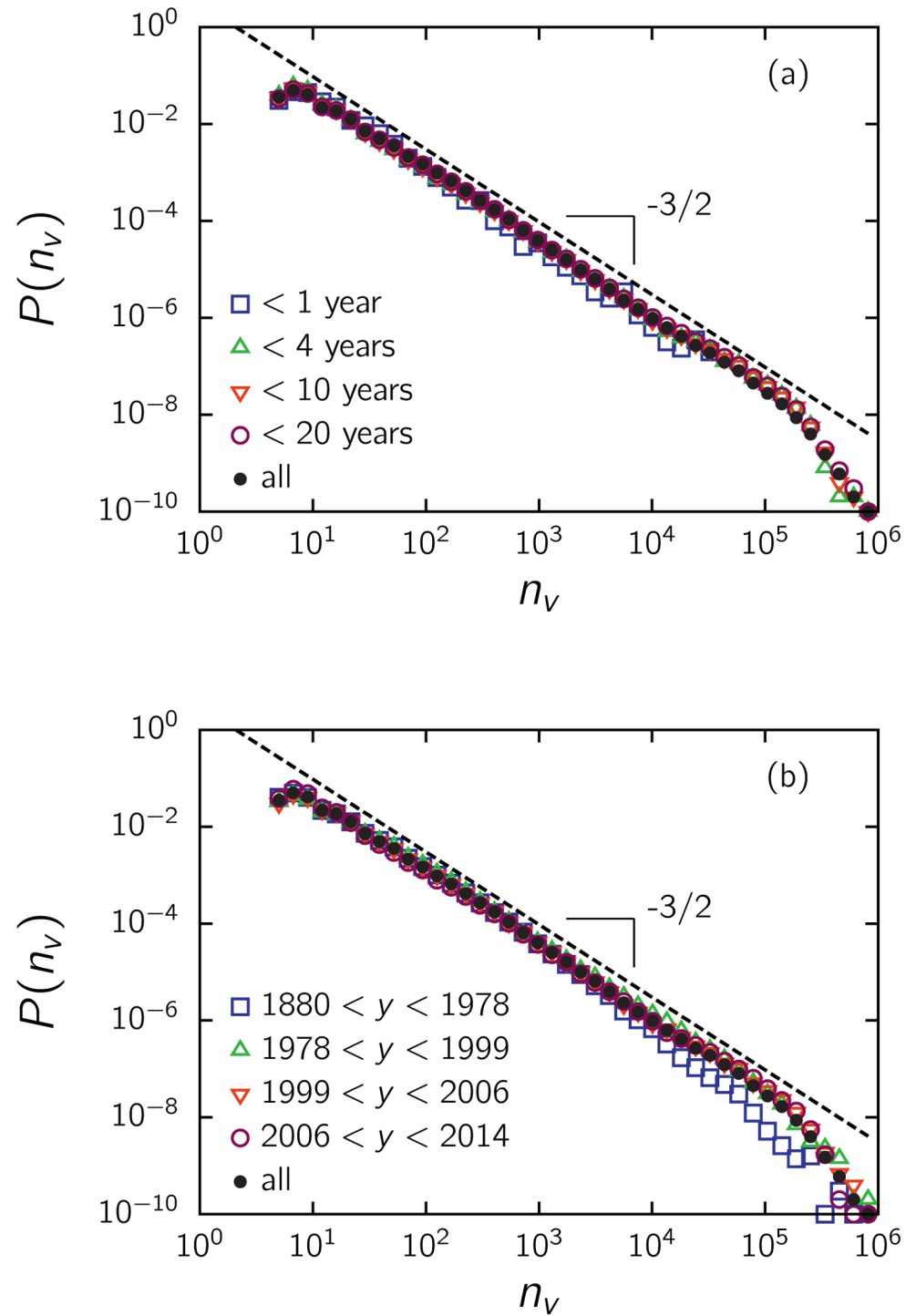


Fig 5. Impact of a movie's age on the distribution of votes. $P(n_v)$ for IMDb movies (a) with less than a given number of years, and (b) released within the interval indicated on the figure, chosen to contain the same number of movies.

doi:10.1371/journal.pone.0136083.g005

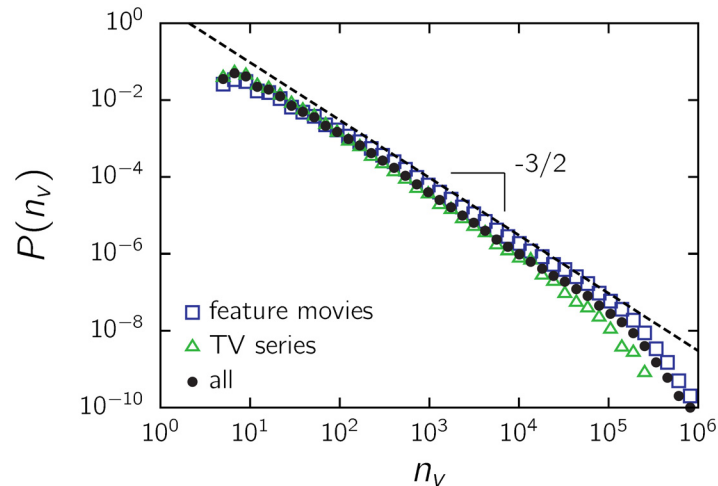


Fig 6. $P(n_v)$ for TV series and feature movies.

doi:10.1371/journal.pone.0136083.g006

($\alpha \approx 1.9$) films, which constitute relatively large subsets, present a noticeable discrepancy, with an exponent close to 2. The audiences of these films appear to have a differentiated response.

Budget

Last but not least, we investigated the dependency between the number of votes and the production budget b_i of each movie. Information about budgets is available for a reduced set of feature films, and we considered those with a budget above 10^3 US\$ (approximately 18 thousand films). For this set, we plotted the number of votes vs. budget (Fig 8). This plot shows that points are scattered but display a positive correlation beyond the first quartile, indicating that above a tipping point, on average, the number of votes increases with the budget, although there exist high-budget films with low appeal and low-budget ones with a moderate response.

The distribution of votes for this set (with budget information) does not have the $3/2$ power law decay that is characteristic of the entire dataset and shown in previous figures (Fig 9a). To analyze this issue more deeply, we divided the dataset using the median and observed that the high-budget half is responsible for that deviation, while the low-budget half preserves the $3/2$ power law. Next, we proceeded to analyze other quantiles. In Fig 9b (and in Fig 8), we see that beyond the level of the median (where the curve given by a non-parametric regression takes off), the distribution completely loses a scaling region. Furthermore, the probability of a large n_v increases with the budget and even develops a peak, as observed for the last percentile. That is, a movie's production budget practically becomes a determinant of the average number of votes it receives. This result shows that a different generative mechanism governs the distribution of votes given to high-budget films, which is not a surprising result because huge budgets include advertising and publicity actions to reach large audiences. The effect of budget may explain why some genres that are typically associated with high production costs, such as action movies and thrillers, present a slightly smaller exponent than the majority of movies.

Modeling

People may select movies based on genre, theme, a cast of actors, directors, producers, etc. They are certainly also influenced by publicity and advertising, which are stronger for high-budget movies; however, these movies will be set aside for the moment in the following discussion.

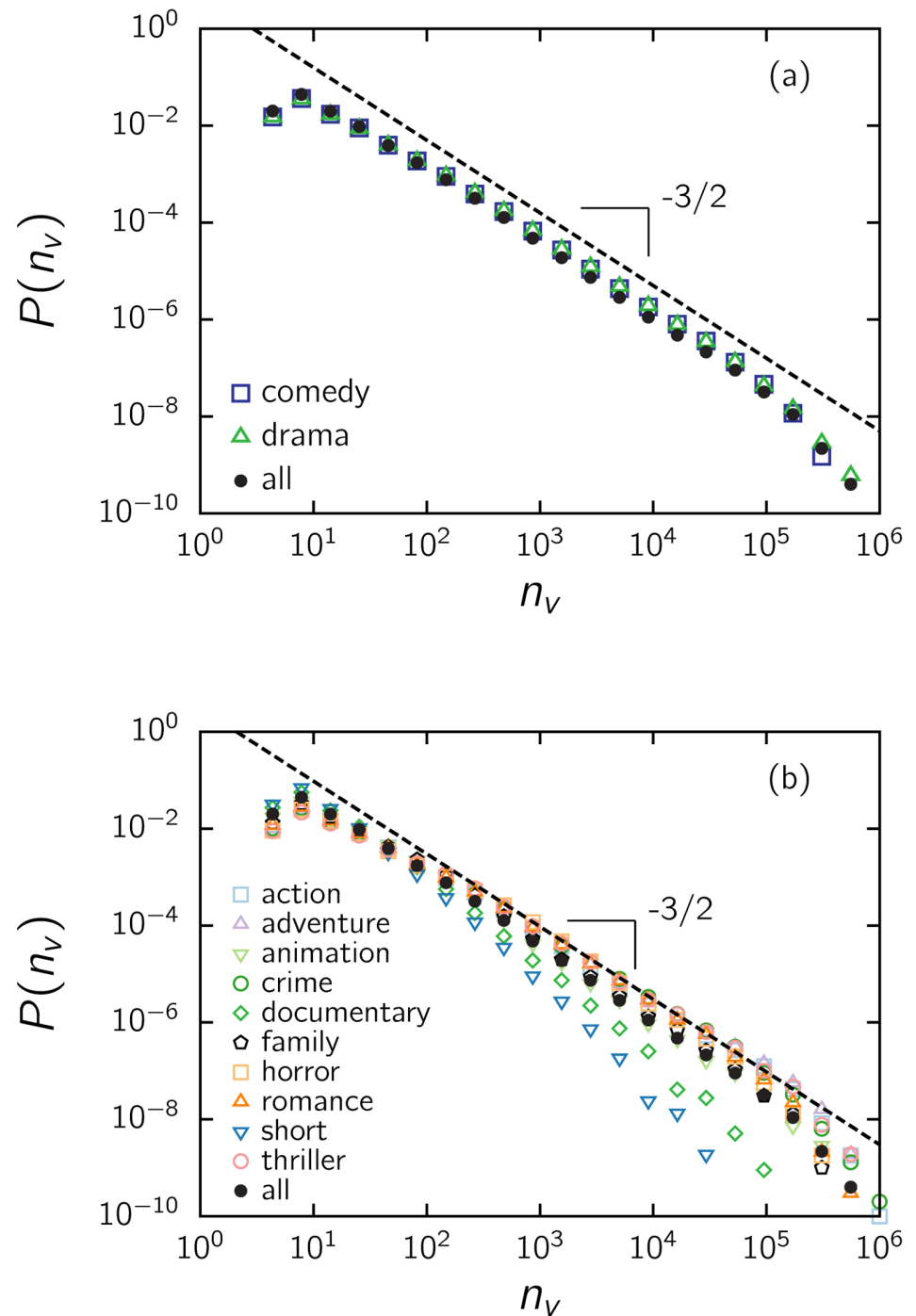


Fig 7. Impact of film genre on the distribution of votes $P(n_v)$ for (a) dramas and comedies and (b) other genres. Some movies belong to more than one genre.

doi:10.1371/journal.pone.0136083.g007

New movies are constantly being released, with actors and directors that are not always sufficiently recognized or popular, and synopses are not always enough to help people decide whether to see a movie or not. Therefore, in many cases, it becomes more practical to adopt other peoples' opinions [16], through recommendation systems, or suggestions of friends and

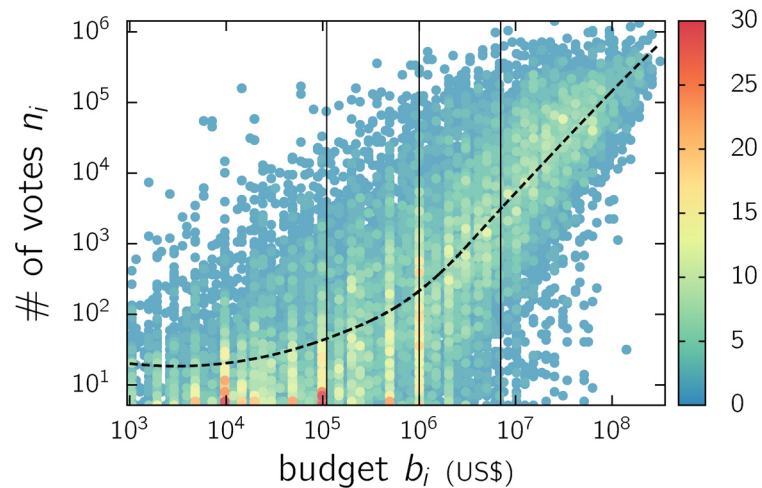


Fig 8. Color map of the number of votes n_i vs. the budget b_i of each IMDb movie i . Each bullet contains the number of movies indicated by the color scale. The vertical lines indicate the quartiles. The dashed line was obtained by means of a non-parametric regression rLOESS [15].

doi:10.1371/journal.pone.0136083.g008

colleagues, before going to the movies or renting a film. It is common for people choose a movie that someone close to them has watched and commented on, also to avoid feeling excluded in ordinary conversations (fear of missing out [17]). In a more general context, it is known that as more people adopt an item, it becomes more likely that somebody else will want to adopt it [16]. Imitation is a common process in many social scenarios and is very useful as a decision strategy. In fact, in Fig 10a, we observe that the increment $\Delta n_v = n_v(t_2) - n_v(t_1)$ for a given time interval $\Delta t = t_2 - t_1$ increases with n_v , indicating some type of cumulative advantage.

Because, individuals must decide between two alternatives, to watch or not watch a movie, their state can be characterized by a binary variable, e.g., active or inactive, as considered in many models of opinion contagion [4, 18–21]. To define the rules of contagion, recall that the empirical value of α is very close to the standard mean-field value of 3/2 for the avalanche size distribution, as in self-organized criticality with random neighbors [22–24], threshold activation [4] and bootstrap percolation in random networks with finite second moment of the degrees [5–7, 25]. The same scaling also occurs for the cluster size distribution of random networks at percolation [26, 27]. Therefore, the mean-field value suggests some randomness in the dissemination process, independently of the precise pattern of contacts. The present empirical findings indicate that typically, the rules of contagion do not need to incorporate intrinsic features of the movies or of particular audiences, although, as a counterexample, the audiences of short and documentary films appear to behave differently.

Some dissemination rules proposed in the literature require exposure of a node to several active nodes for activation, such as in the threshold model developed by Watts [4], where a minimal fraction of consensus among neighbors is required to infect a node. However, movies can currently be accessed through diverse media, and the choice of one movie does not exclude others; therefore, we believe that in the present case, a single enthusiastic contact may be enough to induce the decision to watch a movie. Therefore, we can consider the simplest case in which a single active node is capable of activating (or infecting) some of its neighbors.

IMDb users represent a sample of the population that watches movies, and, as such, user opinions expressed through the rating system are an indicator of the opinions of the general audience. In particular, the number of votes for a movie is a measure of the general audience’s

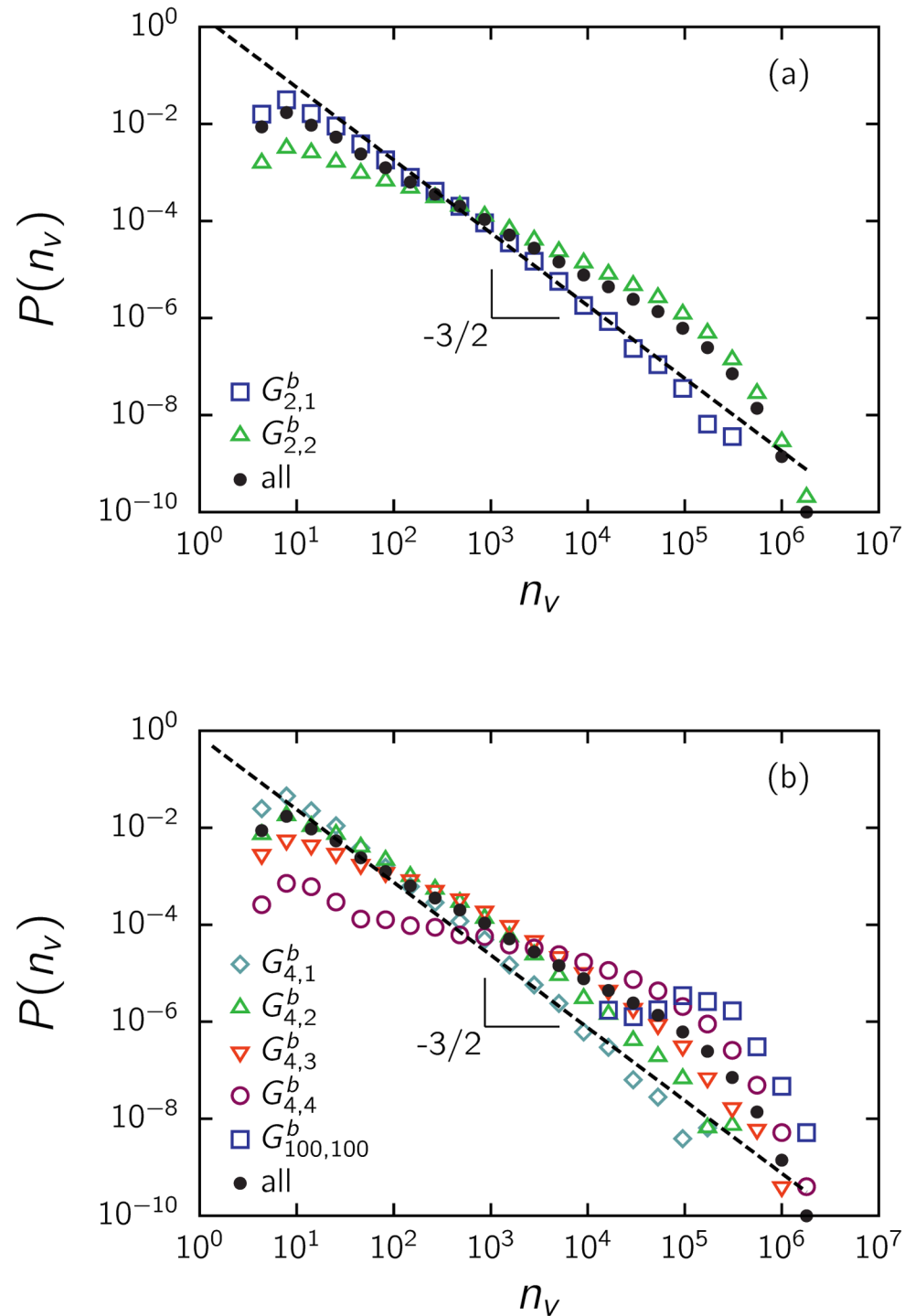


Fig 9. Impact of the budget of feature films on the distribution of votes $P(n_v)$ for (a) the two groups $G_{2,1}^b$ and $G_{2,2}^b$ separated by the median with respect to the budget and (b) the four groups $G_{4,1}^b, \dots, G_{4,4}^b$ using the quartiles, indicated in Fig 8, and the last percentile $G_{100,100}^b$. The dashed line with slope $-3/2$ is drawn for comparison, as well as the distribution for all films with budget information (only feature films with $b_i \geq 10^3$ US\$ were considered).

doi:10.1371/journal.pone.0136083.g009

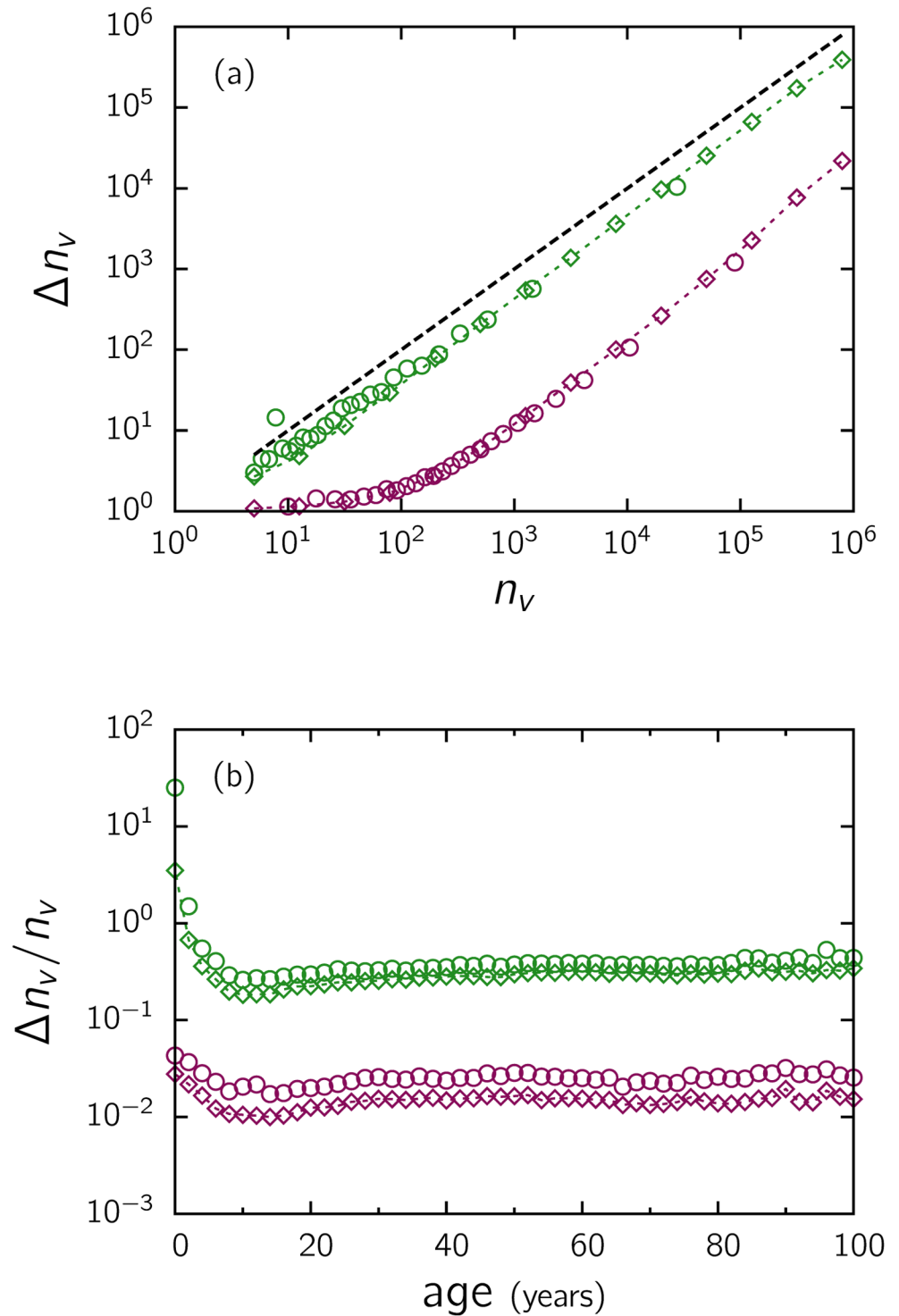


Fig 10. (a) Increment of the number of votes $\Delta n_v = n_v(t_2) - n_v(t_1)$ as a function of n_v and (b) the relative increment $\Delta n_v/n_v = [n_v(t_2) - n_v(t_1)]/n_v(t_1)$ as a function of the age of the movie for $\Delta t = t_2 - t_1 \approx 1$ month (red, obtained with sets 2 and 3) and 22 months (green, obtained with sets 1 and 2). See (S1 Dataset). The same list of movies, with at least 5 votes at t_1 , was considered. The symbols represent the arithmetic (circles) and geometric (diamond) mean values. The dotted lines are a guide; the dashed line in panel (a) with slope 1 was drawn for comparison.

doi:10.1371/journal.pone.0136083.g010

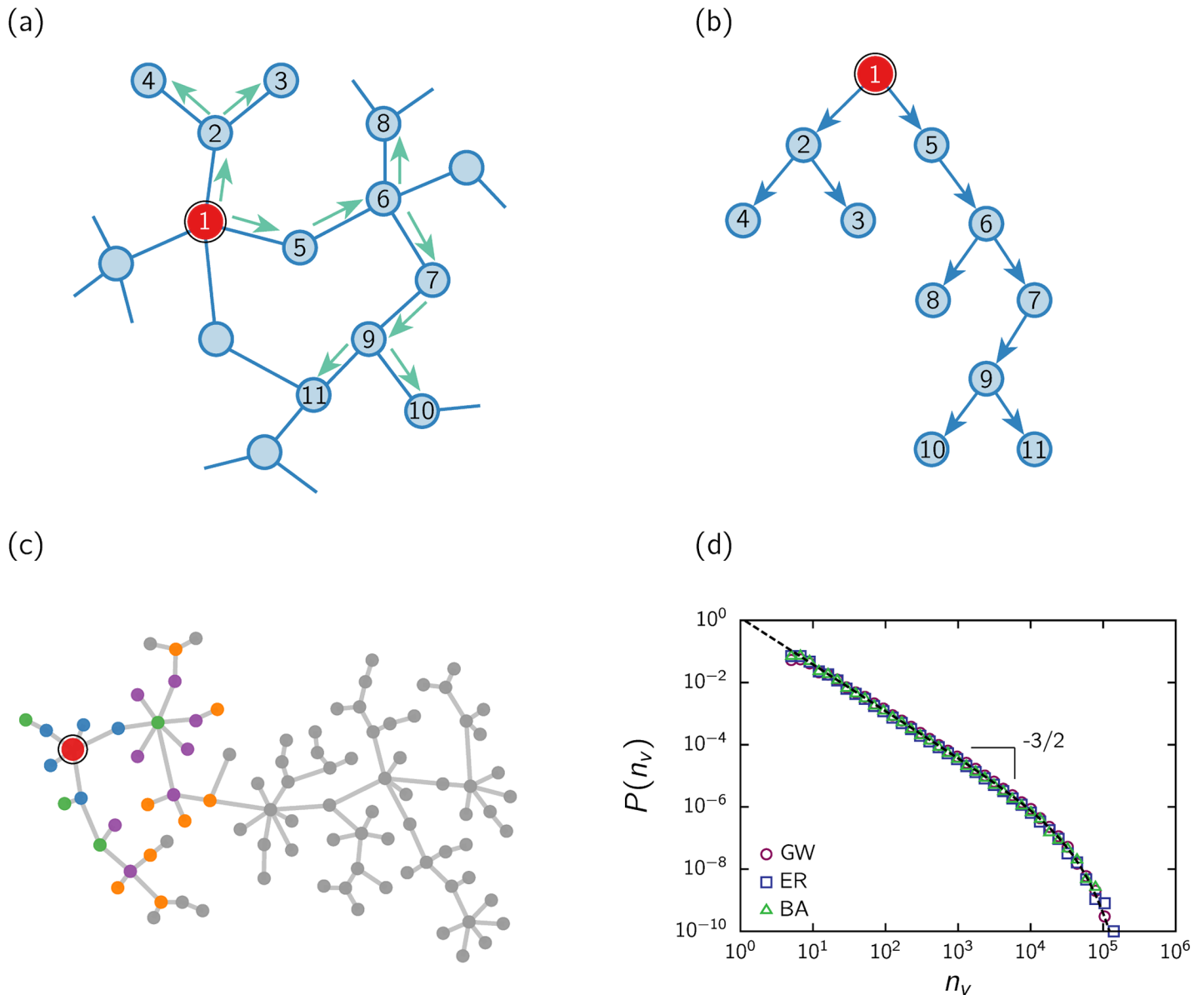


Fig 11. Pictorial representation of the contagion process and the equivalent branching process. (a) Underlying network of contacts. The contagion starts at an initiator node (largest node). Contagion occurs (green arrows) to some of its neighbors (a number of them that we assume to be a random variable) and so on an avalanche develops. (b) A branching tree is built from the contacts that participate of the contagion process. (c) Branching tree realization of a simple Galton-Watson process. The largest node represents the initiator, the first successive generations of the tree are identified with colors, and the final tree is shown as a result of a cascade that becomes extinct at the 13th generation. (d) Distribution of avalanche sizes from simulations of the contagion process: for a simple (network-free) Galton-Watson (GW) process and for the equivalent contagion process on top of Erdős-Rényi (ER) and Barabási-Albert (BA) networks of size 10^6 and average connectivity $\langle k \rangle = 100$. In all cases the probability p_j of influencing j individuals was arbitrarily chosen to be exponential with mean $p \lesssim 1.0$, and 10^6 realizations were considered.

doi:10.1371/journal.pone.0136083.g011

level of interest in that movie. Therefore, we will make the reasonable assumption that the number of votes is proportional to the number of people that watched the movie.

The full *rating process*, that involves giving a score, is a more complicated process than simply voting. The individual opinion about a movie, recorded in the number of stars, is certainly influenced by social interactions and personal preferences. However, based on the empirical evidences, we can model the statistics of the number of people that became interested in

watching a movie, which is manifested in the number of votes, regardless of the scores. This simplification of the full process is motivated by the observation that the main characteristics of the statistics of the number of votes are independent of the score, as well as of other attributes of the movies.

In our modeling, one or a few active initiators propagate the idea of watching a given movie, convincing (or infecting) some of their contacts who, in turn, can infect others, and so on. That is, amongst the k contacts of an activated node, a random number j of them, chosen with probability distribution $\{p_j\}$, with $0 \leq j \leq k < \infty$, becomes activated (see Fig 11a, where we depicted a small connectivity network only for the sake of clearness). The dissemination process stops when, in a given time step, no new nodes become activated. The total number of activated nodes, which we will refer to as an *avalanche* or *cascade*, reflects the number of people who decided to watch the movie.

We simulated this simple contagion process on Erdős-Rényi (ER) [28] and Barabási-Albert (BA) [29] networks, which are representative networks of contacts with homogeneous and scale-free degree distributions, respectively. We assume that the average connectivity $\langle k \rangle$ is large, like in online social networks (e.g., Twitter, Facebook, etc.) or in real-world relationships. Fig 11d, shows the distribution of avalanche sizes resulting from the simulations, using a discrete exponential probability distribution $\{p_j\}$, with mean value $p \lesssim 1$ (The discrete exponential distribution $P(n) = (1 - e^{-\lambda})e^{-\lambda n}$ was used, with $\lambda = 0.7$, hence $p \equiv \langle n \rangle = e^{-\lambda}/(1 - e^{-\lambda}) \simeq 0.986$). The $3/2$ scaling of real data is well reproduced by the model, independent of the network used, as far as the mean connectivity is large enough ($\langle k \rangle \gg p$).

The contagion cascade can be *represented* by a branching process: each active individual is represented by a node, and if that individual convinces a contact, then it generates a new branch of the tree; the number j of new branches of each node is drawn with the probability distribution $\{p_j\}$, independent of the process history and of other branches. At a given generation k of the branching process, there is a total number of nodes m_k . This protocol describes a Galton-Watson (GW) branching process [30]. An avalanche of contagion develops and stops with a certain probability of extinction. Extinction is sure if the mean value $p < 1.0$, otherwise, there is a positive probability of surviving indefinitely. The growth of a tree is analogous to the development of an avalanche in the network. This mapping is illustrated in Fig 11a and 11b. Each movie develops its own independent tree, whose number of nodes reflects the number of people who decided to watch the movie. In Fig 11c, we show a realization in the particular case of the exponential probability distribution $\{p_j\}$.

The total size of the avalanche $n = \sum_{k \geq 1} m_k$ is known to have a distribution $P(n) \propto n^{-3/2}$ with an exponential cutoff [30], as depicted in Fig 11d. The cutoff depends on the details of the distribution $\{p_j\}$, which was chosen in the examples to be exponential, but the exponent $3/2$ is a robust result as long as $\{p_j\}$ meets some minimal requirements [30].

The mapping of the network spreading into the GW branching involves the simplification that interferences or overlaps in the network spreading can be ignored, since, in the GW process, branches are independent. When implemented in a network, as the number of infected nodes increases, the contagion probability should decrease, since, in principle, the number of nodes that could be activated diminishes. However, the approximation that the probability of contagion is constant over time is reasonable when the mean connectivity is large, as assumed ($\langle k \rangle = 100$ in the case of the example). Hence, the distribution of avalanche sizes presents the same scaling, as observed in Fig 11d. Moreover, the outcomes seem independent of the kind of network, which is consistent with the mean-field character of the process.

Despite the simplicity of the model (and the approximations made), it allows to capture the main features of empirical data. It is unnecessary to make any specific assumption about the underlying network of contacts, except that the amount of contacts is large enough. Further

assumptions are irrelevant to reproduce the $3/2$ scaling. However, the categories of short and documentary are counterexamples, displaying a different scaling. In those cases, where the audience is constituted by a singular community, we should have to incorporate more ingredients to the model as will be discussed below.

The type of branching process illustrated in [Fig 11](#) directly maps other models, in particular, the annealed random neighbor version of the sandpile model, a paradigm of self-organized criticality [22–24]. In the random sandpile model, an initiator site topples to its neighbors and a new branch of a tree can be drawn when a neighbor in turn topples, continuing the cascade. In this way, a tree is generated. In that sandpile model, neighbors are randomly chosen at each step so that there is no underlying network of contacts. Even if there is a network of contacts, there are situations, such as when the critical dimensionality is exceeded, where the mean-field character holds. For instance, the mean-field exponent $3/2$ is preserved for sandpile models on exponential-type and scale-free networks with a finite second moment. But the exponent increases for very heterogeneous degree distributions [31–34]. Also, for instance, when there is branch overlapping in the dissemination process, the structure of the network of contacts might become important and hence the associated branching process becomes correlated and concomitantly the mean-field character is lost. Something similar may happen for certain audiences, where a different scaling is observed, such as in the case of short and documentary films. In those cases, we should probably need to take into the consideration details of the networks of contacts and/or correlations in the propagation process that can not be neglected. However, such empirical information is not available.

Typical real adoption curves (the number of adoptions as a function of time) for general items, particularly online movies [35], are known to first experience a pre-takeoff stage after release, then grow rapidly in the early stages, reach a peak and finally decay. The integral of an adoption curve gives the size of the cascade. Usually, a real cascade does not become completely extinct, differently to the simulations in [Fig 11](#), but remains at a low level after a certain time because audiences continue to watch old movies. This effect is shown in [Fig 10b](#), where we plot the relative increment $\Delta n_v/n_v = [n_v(t_2) - n_v(t_1)]/n_v(t_1)$ of the number of votes as a function of the age of the movie (time elapsed since release) for two different time intervals $\Delta t = t_2 - t_1$. In both cases, while the increment is large in the initial stage, after a few years, it decays to a low average level.

High-budget films, not considered till now, produce distinct statistics. High budgets imply not only attractiveness stemming from famous actors and expensive settings but also high publicity and advertising budgets. These play the role of an external field that helps to trigger very large cascades. Therefore, it is not surprising that in these cases, the spread mechanism does not only rely on people's interactions. Instead, without high budgets, interactions could become dominant, and self-organization emerges. Even in this latter case, the power-law scaling indicates that extremely large avalanches are possible, giving rise to blockbusters even among low-budget productions.

Discussion

A long-tailed distribution of votes, with a power law exponent close to $3/2$, emerges from the statistical analysis of IMDb votes. This is a robust result that holds for various subsets defined by movie attributes such as average rating, age or genre, but not for high-budget films. Furthermore, it does not depend on the target audience, suggesting that the dissemination process occurs independently of the pattern of contacts. Genre exceptions are shorts and documentary films, whose audiences show a less heterogeneous pattern.

Remarkably, regardless of whether a movie is “good” enough to be well rated, the scale-free character of the statistics is not affected. The universality of the results indicates that modeling

does not require accounting for intrinsic features of the movies or of particular audiences. A type of process that can be naturally associated to the underlying propagation of adoptions is a random multiplicative or branching process, whereby activations (or adoptions) multiply, randomly beginning from a few initial adopters, as illustrated in Fig 11. In this uncorrelated case, cascades emerge with a size distribution decaying as a power law with exponent $3/2$. In fact, branching processes are the skeleton of many activation models, producing avalanches in the same universality class [4–7, 25]. As a consequence, the empirical outcomes fit a scenario of imitation cascades. Assuming that voters are a statistically significant sample of the audience, our results represent empirical evidence of this type of cascade.

Meanwhile, high-budget films yield very different statistics, which plausibly reflects that the marketing campaigns associated with large budgets overcome interpersonal activity, acting as an external field that destroys the scaling behavior. Therefore, although the audience—not a movie’s characteristics—makes a movie a hit or a blockbuster [36], large enough budgets can overcome this natural trend.

To evaluate the extent to which IMDb users constitute a biased sample of movie audiences, we considered another similar source, Netflix [9]. We observed the same $3/2$ scaling, even though this database contains primarily commercially attractive films, while IMDb is more broad.

The possibility of flaws in the rating system, such as blind voting or massive fan voting, or imprecision in the provided information, cannot be discarded; however, isolated cases are not expected to influence the statistics. Also IMDb is not a static database; the information is constantly being refined by people adding and correcting the data through their “Contributor Zone” section.

Whether other items (consumer goods or cultural products) which can be rated online follow similar propagation processes remains to be investigated. Unlike IMDb, other websites allowing user reviews, e.g., the Apple App Store and Google Play, sell the products being rated, which may mold choices differently. People also choose “items” in political electoral processes. However, as far as the distribution of political votes follows log-normal or power law patterns with different exponents [37–39], the propagation of decisions in that context is of a very different nature because the correlations in the interaction network seem to be relevant. In contrast, in the current case, the dissemination process appears to be uncorrelated. This randomness may be at the root of the difficulties in optimizing recommender systems [3].

Beyond providing a complete explanation for the complex phenomenology related to the spread of adoptions, in this work, we aimed to expand the sources of empirical information related to the field of social dynamics. It is worth remembering that much of the progress in the complex networks field has been driven by empirical observations of real networks. The tools available to rate products can reveal some interesting properties of online social systems, such as the patterns presented in this work.

Supporting Information

S1 Dataset. Description of data from IMDb. We collected votes (from 1 to 10 stars) for all movies, excluding TV episodes (total number of 336,090,882 votes for 300,723 movies), from March 19 to 28, 2013 (set # 1). Using the same list of movies, we collected the number of votes again from December 8 to 18, 2014 (set #2, 465,292,451 votes) and from January 5 to 10, 2015 (set # 3, 471,222,420), as shown in (Fig 10). For budgets, we use a new list and collected data from February 5 to 8, 2015. Results with fewer than 5 votes (in 2013) are not exhibited. Number of items by type: 33,941 (Documentary) 133,775 (Feature Film) 3,172 (Mini-Series) 50,408 (Short Film) 1,071 (TV Episode) 25,168 (TV Movie) 33,165 (TV Series) 2,450 (TV Special)

12,120 (Video) 5,453 (Video Game) By genre: 24,911 (Action); 93 (Adult); 15,651 (Adventure); 18,918 (Animation); 5,385 (Biography); 74,393 (Comedy); 18,693 (Crime); 37,250 (Documentary); 97,087 (Drama); 16,022 (Family); 8,677 (Fantasy); 567 (Film Noir); 1,575 (Game Show); 5,525 (History); 15,072 (Horror); 10,212 (Music); 5,840 (Musical); 8,170 (Mystery); 1,036 (News); 3,605 (Reality TV); 21,165 (Romance); 8,239 (Sci-Fi); 61,538 (Short); 4,360 (Sport); 1,467 (Talk Show); 16,246 (Thriller); 5,080 (War); 4,549 (Western). An item could be defined by more than one genre. As a final observation, it is possible for a user to remove his or her vote; as a consequence, a small fraction of movies have a decreasing number of votes. However, this represents a negligible fraction of the movies. We used the following list: http://www.imdb.com/search/title?title_type=feature,tv_movie,tv_series,tv_special,mini_series,documentary,game,short,video,unknown&user_rating=1.0,10.

(ZIP)

Acknowledgments

We acknowledge financial support from Brazilian Agencies CNPq, CAPES and Faperj.

Author Contributions

Analyzed the data: MR AMC CA. Wrote the paper: MR AMC CA.

References

- Castellano C, Fortunato S, Loreto V. Statistical physics of social dynamics. *Reviews of Modern Physics*. 2009; 81(2):591–646.
- Galam S. *Sociophysics: a physicist's modeling of psycho-political phenomena*. Springer Science & Business Media; 2012.
- Lü L, Medo M, Yeung CH, Zhang YC, Zhang ZK, Zhou T. Recommender systems. *Physics Reports*. 2012; 519(1):1–49. doi: [10.1016/j.physrep.2012.02.006](https://doi.org/10.1016/j.physrep.2012.02.006)
- Watts DJ. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*. 2002; 99(9):5766–5771. doi: [10.1073/pnas.082090499](https://doi.org/10.1073/pnas.082090499)
- Baxter GJ, Dorogovtsev SN, Goltsev AV, Mendes JFF. Bootstrap percolation on complex networks. *Physical Review E*. 2010; 82(1):011103. doi: [10.1103/PhysRevE.82.011103](https://doi.org/10.1103/PhysRevE.82.011103)
- Baxter GJ, Dorogovtsev SN, Goltsev AV, Mendes JFF. Heterogeneous k-core versus bootstrap percolation on complex networks. *Physical Review E*. 2011; 83(5):051134. doi: [10.1103/PhysRevE.83.051134](https://doi.org/10.1103/PhysRevE.83.051134)
- Goltsev AV, Dorogovtsev SN, Mendes JFF. k-core (bootstrap) percolation on complex networks: Critical phenomena and nonlocal effects. *Physical Review E*. 2006; 73(5):056101. doi: [10.1103/PhysRevE.73.056101](https://doi.org/10.1103/PhysRevE.73.056101)
- <http://www.similarweb.com/website/imdb.com>; <http://www.alexa.com/topsites>.
- <http://www.netflixprize.com>.
- Newman MEJ. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*. 2005; 46(5):323–351. doi: [10.1080/00107510500052444](https://doi.org/10.1080/00107510500052444)
- Mitzenmacher M. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*. 2004; 1(2):226–251. doi: [10.1080/15427951.2004.10129088](https://doi.org/10.1080/15427951.2004.10129088)
- Koenigstein N, Dror G, Koren Y. Yahoo! music recommendations: modeling music ratings with temporal dynamics and item taxonomy. In: *Proceedings of the fifth ACM conference on Recommender systems*. ACM; 2011. p. 165–172.
- Lorenz J. Universality in movie rating distributions. *The European Physical Journal B*. 2009; 71(2):251–258. doi: [10.1140/epjb/e2009-00283-3](https://doi.org/10.1140/epjb/e2009-00283-3)
- <http://www.imdb.com/features/15thanniversary/history>.
- Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*. 1979; 74(368):829–836. doi: [10.1080/01621459.1979.10481038](https://doi.org/10.1080/01621459.1979.10481038)

16. Salganik MJ, Watts DJ. Web-Based Experiments for the Study of Collective Social Dynamics in Cultural Markets. *Topics in Cognitive Science*. 2009; 1(3):439–468. doi: [10.1111/j.1756-8765.2009.01030.x](https://doi.org/10.1111/j.1756-8765.2009.01030.x) PMID: [25164996](https://pubmed.ncbi.nlm.nih.gov/25164996/)
17. Przybylski AK, Murayama K, DeHaan CR, Gladwell V. Motivational, emotional, and behavioral correlates of fear of missing out. *Computers in Human Behavior*. 2013; 29(4):1841–1848. doi: [10.1016/j.chb.2013.02.014](https://doi.org/10.1016/j.chb.2013.02.014)
18. de Oliveira MJ, Mendes JFF, Santos MA. Nonequilibrium spin models with Ising universal behaviour. *Journal of Physics A: Mathematical and General*. 1993; 26(10):2317. doi: [10.1088/0305-4470/26/10/006](https://doi.org/10.1088/0305-4470/26/10/006)
19. Galam S. Majority rule, hierarchical structures, and democratic totalitarianism: A statistical approach. *Journal of Mathematical Psychology*. 1986; 30(4):426–434. doi: [10.1016/0022-2496\(86\)90019-2](https://doi.org/10.1016/0022-2496(86)90019-2)
20. Chen P, Redner S. Consensus formation in multi-state majority and plurality models. *Journal of Physics A: Mathematical and General*. 2005; 38(33):7239. doi: [10.1088/0305-4470/38/33/003](https://doi.org/10.1088/0305-4470/38/33/003)
21. Shao J, Havlin S, Stanley HE. Dynamic opinion model and invasion percolation. *Physical Review Letters*. 2009; 103(1):018701. doi: [10.1103/PhysRevLett.103.018701](https://doi.org/10.1103/PhysRevLett.103.018701) PMID: [19659181](https://pubmed.ncbi.nlm.nih.gov/19659181/)
22. Alstrøm P. Mean-field exponents for self-organized critical phenomena. *Physical Review A*. 1988; 38(9):4905. doi: [10.1103/PhysRevA.38.4905](https://doi.org/10.1103/PhysRevA.38.4905) PMID: [9900967](https://pubmed.ncbi.nlm.nih.gov/9900967/)
23. Christensen K. Self-organization in models of sandpiles, earthquakes, and flashing fireflies. *Institute of Physics and Astronomy*; 1992.
24. Bak P. *How nature works: the science of self-organized criticality*. Springer Science & Business Media; 2013.
25. Ramos M, Shao J, Reis SDS, Anteneodo C, Andrade JS Jr, Havlin S, et al. How does public opinion become extreme? *Scientific Reports*. 2015; 2:10032. doi: [10.1038/srep10032](https://doi.org/10.1038/srep10032)
26. Newman MEJ, Strogatz SH, Watts DJ. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*. 2001; 64(2):026118. doi: [10.1103/PhysRevE.64.026118](https://doi.org/10.1103/PhysRevE.64.026118)
27. Centola D, Eguiluz VM, Macy MW. Cascade dynamics of complex propagation. *Physica A: Statistical Mechanics and its Applications*. 2007; 374(1):449–456. doi: [10.1016/j.physa.2006.06.018](https://doi.org/10.1016/j.physa.2006.06.018)
28. Erdős P, Rényi A. On random graphs I. *Publ Math Debrecen*. 1959; 6:290–297.
29. Barabási AL, Albert R. Emergence of scaling in random networks. *Science*. 1999; 286(5439):509–512. doi: [10.1126/science.286.5439.509](https://doi.org/10.1126/science.286.5439.509) PMID: [10521342](https://pubmed.ncbi.nlm.nih.gov/10521342/)
30. Harris TE. *The theory of branching processes*. Springer-Verlag, Berlin; 1963.
31. Goh KI, Lee DS, Kahng B, Kim D. Sandpile on scale-free networks. *Physical Review Letters*. 2003; 91(14):148701. doi: [10.1103/PhysRevLett.91.148701](https://doi.org/10.1103/PhysRevLett.91.148701) PMID: [14611564](https://pubmed.ncbi.nlm.nih.gov/14611564/)
32. Lee DS, Goh KI, Kahng B, Kim D. Branching process approach to avalanche dynamics on complex networks. *Journal Korean Physical Society*. 2004; 44(1):633–637.
33. Bonabeau E. Sandpile dynamics on random graphs. *Journal of the Physical Society of Japan*. 1995; 64:327. doi: [10.1143/JPSJ.64.327](https://doi.org/10.1143/JPSJ.64.327)
34. Olami Z, Christensen K. Temporal correlations, universality, and multifractality in a spring-block model of earthquakes. *Physical Review A*. 1992; 46(4):R1720–R1723. doi: [10.1103/PhysRevA.46.R1720](https://doi.org/10.1103/PhysRevA.46.R1720) PMID: [9908369](https://pubmed.ncbi.nlm.nih.gov/9908369/)
35. Dover Y, Goldenberg J, Shapira D. Network traces on penetration: Uncovering degree distribution from adoption data. *Marketing Science*. 2012; 31(4):689–712. doi: [10.1287/mksc.1120.0711](https://doi.org/10.1287/mksc.1120.0711)
36. De Vany A, Walls WD. Uncertainty in the movie industry: Does star power reduce the terror of the box office? *Journal of Cultural Economics*. 1999; 23(4):285–318.
37. Fortunato S, Castellano C. Scaling and universality in proportional elections. *Physical Review Letters*. 2007; 99(13):138701. doi: [10.1103/PhysRevLett.99.138701](https://doi.org/10.1103/PhysRevLett.99.138701) PMID: [17930647](https://pubmed.ncbi.nlm.nih.gov/17930647/)
38. Costa Filho RN, Almeida MP, Andrade JS, Moreira JE, et al. Scaling behavior in a proportional voting process. *Physical Review E*. 1999; 60(1):1067. doi: [10.1103/PhysRevE.60.1067](https://doi.org/10.1103/PhysRevE.60.1067)
39. Lyra ML, Costa UMS, Costa Filho RN, Andrade JS Jr. Generalized Zipf's law in proportional voting processes. *EPL (Europhysics Letters)*. 2003; 62(1):131. doi: [10.1209/epl/i2003-00371-6](https://doi.org/10.1209/epl/i2003-00371-6)