SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
APPLIED PHYSICS
COMPUTATIONAL SCIENCE
SCIENTIFIC DATA
APPLIED MATHEMATICS

Received
3 April 2013

Accepted
28 August 2013

Published
26 September 2013

Correspondence and requests for materials should be addressed to S.S. (sreens@rpi.edu)

Quantitative analysis of the evolution of novelty in cinema through crowdsourced keywords

Sameet Sreenivasan[1,2,3]

[1]Social and Cognitive Networks Academic Research Center, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY, 12180, USA, [2]Department of Computer Science, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY, 12180, USA, [3]Department of Physics, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY, 12180, USA.

The generation of novelty is central to any creative endeavor. Novelty generation and the relationship between novelty and individual hedonic value have long been subjects of study in social psychology. However, few studies have utilized large-scale datasets to quantitatively investigate these issues. Here we consider the domain of American cinema and explore these questions using a database of films spanning a 70 year period. We use crowdsourced keywords from the Internet Movie Database as a window into the contents of films, and prescribe novelty scores for each film based on occurrence probabilities of individual keywords and keyword-pairs. These scores provide revealing insights into the dynamics of novelty in cinema. We investigate how novelty influences the revenue generated by a film, and find a relationship that resembles the Wundt-Berlyne curve. We also study the statistics of keyword occurrence and the aggregate distribution of keywords over a 100 year period.

Over the last century, cinema has carved out an indelible niche in human culture, and filmmaking has come to be regarded as an art-form its own right. The film industry of the United States in particular, has had a major influence on the evolution of cinema over the course of its history, and is currently the third largest producer of films in the world, with a global audience and a gross turnover averaging 29.5 billion US dollars over the last five years reported[1]. Despite the fact that trends associated with films, the dissection of their respective successes and failures, and their individual artistic merit are all subjects of avid debate and discussion in the public realm, and although the economics of film has been extensively researched[2], no studies, to our knowledge, have quantitatively analyzed the large scale features of novelty in film plots and the patterns associated with their evolution. With the advent of culturomics as an emerging science[3], it is natural to attempt to bridge this gap with the aid of comprehensive sources of film data such as the Internet Movie Database (IMDb).

The Internet Movie Database (www.imdb.com) is a comprehensive online database containing information on films, television programs and videogames which, according to the site, has "more than 100 million data items including more than 2 million movies". This in large part is made possible by allowing registered users of the site to add new database items or edit the information associated with existing ones. One such category of user-generated information at the center of this study, is that of *plot-keywords* consisting of single words, or word-strings associated with each item. If a keyword proposed by a user is semantically close to a keyword that already exists (i.e., has already been created for association with one or more films), then the user is prompted to use the existing keyword, thus suppressing the creation of synonymous keywords. In the context of films, keywords describe any of a number of aspects of film including but not limited to thematic plot-elements (*father-son-relationship*, *power*, *fame*), specific story elements (*tied-to-a-chair*, *held-at-gunpoint*, *breaking-and-entering*), location references (*manhattan-new-york-city*, *coffee-shop*, *Chevron-gas-station*) specific visual or object references (*life-magazine*, *characters-point-of-view-camera-shot*, *coin-flipping-in-the-air*) or high-level features of the film (*independent-film*, *female-nudity*, *cult-film*). Plot-keywords are thus qualitative descriptors spanning several scales of detail and specificity, and they potentially constitute a rich information set capable of yielding valuable insights into the evolution of films over time.

The dynamics of tagging - the process of users contributing keywords to associate with specific items - as well as folksonomy - the classification of items based on these collective tags - have been widely studied in the context of blogs, photo-sharing and social bookmarking[4–13]. A general consensus derived from these studies is that despite a lack of central control, shared vocabularies with stable probability distributions over words emerge as a result of

collaborative tagging. For example, Halpin et al.[4] showed that the relationship between the frequency of a tag's usage and its rank (based on how frequently it is used) is a power-law, and further proposed a model for tagging dynamics based on preferential attachment that could yield such a relationship. Almost concurrently, Cattuto et al.[5] showed that the frequency-rank plot for tags obtained from *Del. icio. us* and *Connotea* indicated a power-law relationship, and demonstrated that a Yule-Simon model with long-term memory for tagging dynamics could yield this relationship. In the context of information retrieval, Levy and Sandler[9] showed how social tags associated with musical tracks (on a *Last. fm* dataset) defined a semantic space that could enable efficient mood-based clustering and retrieval. Similarly, there have been several studies[10–13] that have focussed on exploring the use of tags for personalized recommendation and query based retrieval. As a representative example, Szomnsor et al.[11], investigated the extent to which combining tags obtained from IMDb and ratings data obtained from Netflix could generate better taste profiles for users, and thus yield a predictor of their ratings for an unseen film. Finally, although unrelated to social tagging but still within the larger domain of collaborative editing, Mestyán et al.[14] showed how user activity data on Wikipedia pertaining to a particular film's entry could yield an early predictor for the box-office success of the film.

In contrast to the above studies, the motivation of this work is to utilize the IMDb plot-keywords dataset as a window into the evolution of films and their content over the course of the last century, and in the process investigate certain aspects of novelty generation in the arts. The characterization of novelty, and the processes that lead to it, have been subjects of thorough investigation in psychology and social science[15–17]. Several of these studies emphasize the role of the combinational process - one that combines existing ideas in a manner not encountered earlier - in novelty generation, in contrast to the process of introducing fundamentally new concepts from scratch. Another aspect of sustained research interest[18–21] is the relationship between the novelty of an item and the hedonic value (or pleasure) derived by an individual upon its consumption. The standard paradigm here, resulting from the pioneering work of Wundt[22] and Berlyne[23], is captured by the Wundt-Berlyne curve, which posits that increasing novelty initially results in increasing hedonic value until it reaches a maximum. Further increasing novelty beyond this intermediate level results in a rapid decline in hedonic value. In summary, the inverted-U shaped Wundt-Berlyne curve posits that individuals seek a balance between familiarity and novelty, shying away from the banal and more strongly from the radically unfamiliar.

The issues of novelty creation and novelty optimization are undoubtedly relevant to the business of cinema. A significant portion of film criticism, commentary and discussion is devoted to analyzing the novelty in the writing and execution of film plots. In addition, one among the various factors responsible in successfully securing the financing and distribution of a film, is its conformity to current trends and past conventions. However, little is known in a quantitative sense regarding the degree to which the competing objectives of novelty and conformity are balanced in the process of new content creation. The plot-keywords dataset has the potential to serve as a starting point in addressing these issues. In addition, it allows us to ascribe novelty scores to films on the basis of their content, including not just elements of the underlying story, but also elements that encapsulate the tone and style of the final finished product. With this goal in mind, we analyze the plot-keywords associated with films produced in the United States over the period between and including the years 1890 and 2011, define two novelty scores based on them, and study the aggregate patterns in novelty evolution over a 70 year period. In addition, we also provide a number of quantitative insights into the probability distribution of plot-keywords over the entire dataset spanning 100 years, and the statistics of their use over time.

Results

We begin by presenting some basic characteristics of the dataset under consideration. Henceforth for brevity, we will refer to plot-keywords simply as "keywords".

Statistics of films and tagging. Figure 1(a) shows the total number of English-language films originating in the US (see Methods for details) each year starting from the earliest recorded entry in the year 1890 through 2011. The number of films produced increases sharply starting around 1907, and corresponds to the "Nickelodeon boom" i.e., the sudden increase in the production of films as a result of the success of the Nickelodeon theater in 1905, which led to the proliferation of theaters devoted to film projection for a mass audience. The majority of the films produced in this period had runtimes of 10–15 minutes[24], and are classified as "Short" under the IMDb-Genre field. To obtain the dataset that forms the core of this study, we considered only feature length films, and additionally only those which were non-adult and non-documentary theatrical releases. As expected the peak around 1910 disappears in the filtered set. Analogous to the Nickelodeon boom, there is a sharp rise in the number of films around the mid-1990s. This is a manifestation of the dramatic increase in independent-film production that occurred in the 1990s and that, by the end of the decade, led to over half the feature length films being produced coming from independent studios and producers[25].

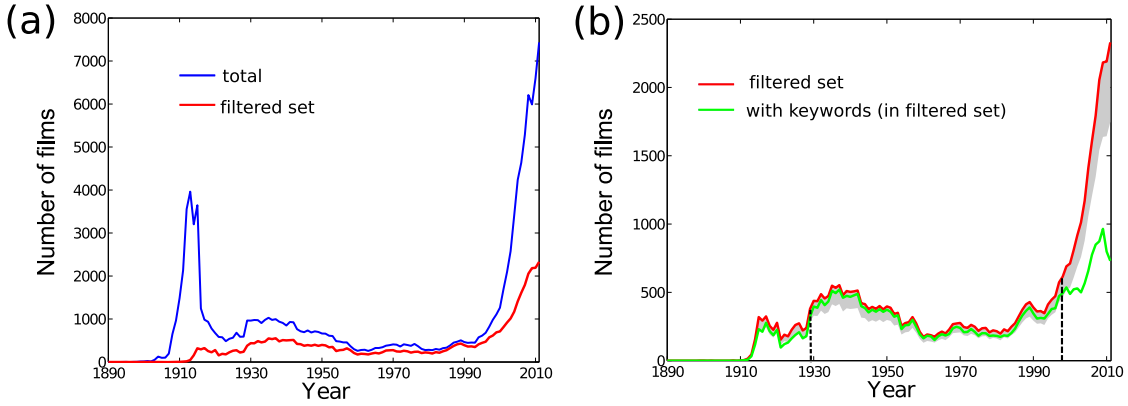

Figure 1 | (a) The total number of English language films produced in the United States (in blue), and the number of films remaining after filtering out short films, documentaries and adult films (in red), per year. (b) Number of films in the filtered set (red) and number of films in the filtered set with keywords (green), per year. The shaded gray region bounds the values which lie within 25% of the total number of films released. In the period between and including the years 1929 and 1998 the green curve lies within the shaded region showing that greater than 75% of films released each year in this period have keywords associated with them.

Figure 1(b) shows the statistics for the tagging of films released in the period between 1890 and 2011. Clearly, the association of keywords to films is not consistent over the different release years, with a clear paucity in tagging towards the early (the first film associated with a keyword was released in 1910) and late years in the period under consideration. However, for years in the period including and between the years 1929 and 1998, more than 75% of the films released each year have keywords associated with them. For our studies on the novelty of films, we therefore focus on the films released within this period. In total, there are 21, 583 films possessing at least one keyword in this period.

We refer to the collection of all keywords associated with a film as the film's *keyword set*. The length of keyword sets appears to be exponentially distributed (see Fig. 2 (a)), with the median length being 14 keywords. For the restricted set between 1929 and 1998, the median length increases slightly to 19, but the distribution remains qualitatively similar (not shown). As expected, films in the tail mostly comprise of popular mainstream films, as shown in Fig. 2(b) for each decade from the 1930s to the 2000s.

Studies on the Google n-gram corpus have demonstrated that trajectories of word-occurrence-frequency over time can reflect surges of cultural interest in specific events, literary works, persons etc.[3,26]. We can expect to glean similar insights from observing the usage of plot-keywords. We begin by defining occurrence frequency per year for a given keyword as the number of films released that year that are tagged with the keyword, divided by the total number of films released that year. Figure 3(a) shows trajectories of occurrence frequency for four example keywords. Similarly as observed for words in literature[3,26], films too display temporally local bursts in the usage of a plot-element as can be seen in the example of "world-war-two". A surge in the occurrence of "class-difference" around 1985 is suggestively coincident with the conjectured rise in materialistic attitudes during the 1980s[27,28].

Beyond the temporally local trends seen in the association of keywords with films, there could also be long-range correlations present. To probe this further, we use the method of detrended fluctuation analysis (DFA)[29] that is widely employed for investigating the presence of long-range correlations in general time-series, and has also been specifically used in the context of word usage[26]. We analyzed using DFA (see Methods), the time series of keyword occurrence frequency for all keywords that appeared in at least 75 of the years between the period 1910 - the earliest year with a tagged film - and 2011. In total, there are 461 such keywords. The Hurst exponent α which signals the presence or absence of long range correlations is obtained for each of

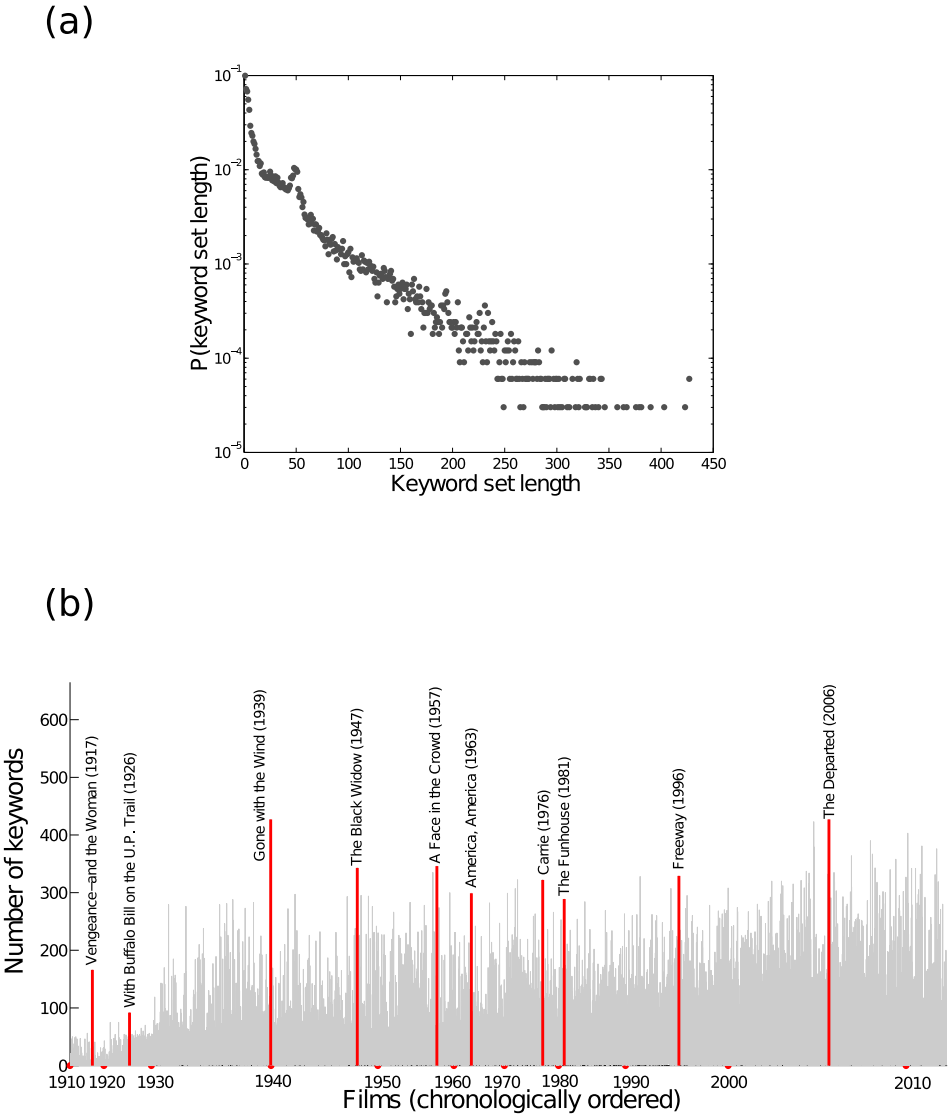

Figure 2 | (a) The distribution of keyword set lengths over all films with keywords. The linear decay on the linear-log plot indicates a roughly exponentially declining probability as the keyword set length increases. (b) Length of the keyword set for the chronologically ordered set of films with keywords. The gray bars indicate the lengths of the sets for the different films. For each decade, the film with the longest keyword set over all releases in that decade is highlighted in red.

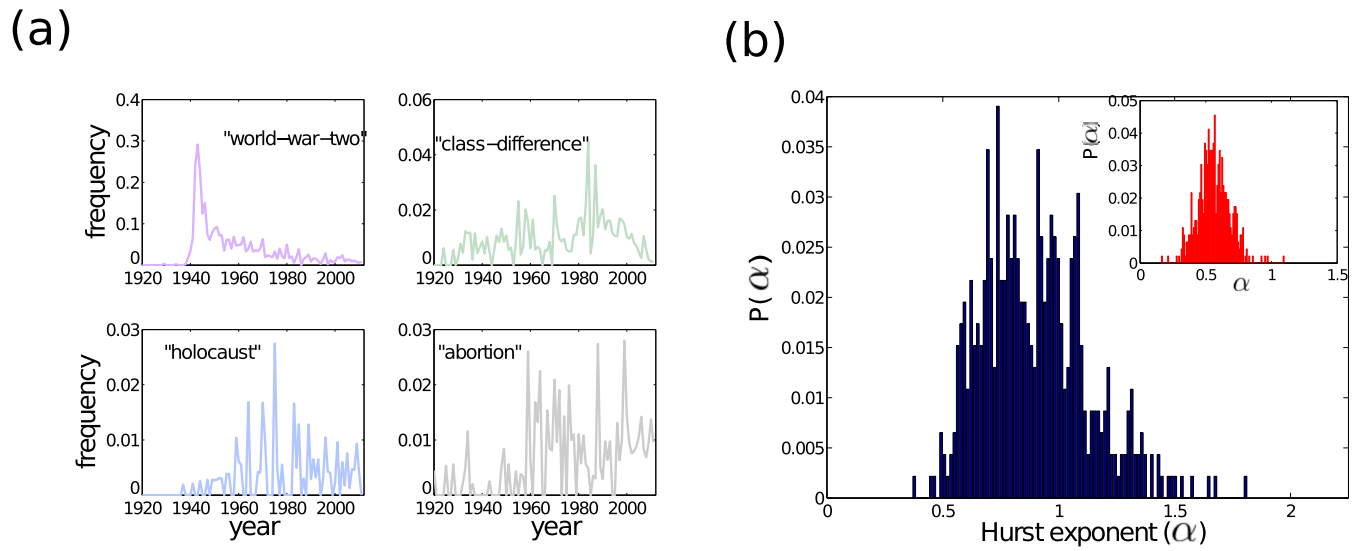

Figure 3 | (a) The yearly occurrence frequency of specific keywords as a function of time. (see text for details)(b) Distribution (relative frequencies) of the Hurst exponent α for keywords that occur in at least 70 years between the period 1910 and 2011. The mean value of the exponent is 0.8966, indicating the presence of positive long-range correlations. The inset shows the distribution after shuffling each of the time series. The correlations largely disappear upon shuffling as indicated by the mean value of 0.5590 obtained for α.

these time series using DFA. A value of $\alpha = 0.5$ indicates no temporal correlations, $\alpha < 0.5$ indicates negative correlations while $\alpha > 0.5$ indicates positive correlations. A distribution of the Hurst exponents obtained for the 461 time series considered is shown in Figure 3(b), indicating the presence of long-range positive correlations in the keyword occurrence frequency. These correlations disappear (see Fig. 3(b) inset) for the set of time series obtained after shuffling the temporal order of data within each individual time series.

Evolution of film novelty. Next, we devise a method to assign a novelty score to each film on the basis of the keywords associated with it and the keywords appearing in all films that were released prior to it. The assignment of novelty scores is done for films in the continuous period between and including 1929 and 1998, more than 75% of which per year are associated with keywords. In addition to the fact that keyword data is abundantly present for films released in or after 1929, the choice of this year as the beginning of our time window is also motivated by the fact that by then, film-going was no longer an esoteric form of entertainment[24] with film-ticket sales in the 1930s constituting as much as 4/5ths of all entertainment expenditure[30]. Incidentally, the year 1929 marks the time around which sound in films became ubiquitous[24], the beginning of the period which came to be known as the golden age of Hollywood[31], and the year in which the first ever academy awards were presented. We formally present the definition of the novelty score below.

For a film i, denote by M^i the set of all films that appear prior to the release of film i. We use m to index an arbitrary film, and K_m to be the set of keywords associated with m. We begin by computing the probability $P(w)$ of observing a keyword w over the set of films $M^i \cup \{i\}$ for all keywords appearing in the set.

$$P(w) = \frac{1}{|M^i|+1} \sum_{m \in M^i \cup \{i\}} \mathbb{1}_{K_m}(w) \tag{1}$$

where $\mathbb{1}_A$ denotes the indicator function for set A:

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \tag{2}$$

Then, for any keyword w, the quantity $-\log P(w)$ is a standard measure of the "surprise" in observing keyword w[32]. With this in mind, we can quantify the novelty of film i, as the average surprise over all keywords associated with the film. Although, ideally, $P(w)$ should designate the prior probability distribution i.e., the probability distribution for keywords computed over films in M^i, we include film i in its computation in order to circumvent the ill-defined logarithm arising when $P(w) = 0$ i.e., when w appears for the first time in K_i. The first measure of novelty we define, aims to score the film on the basis of how rarely, on average, the elements associated with it have appeared in films in the past. For a given film i, the average surprise associated with its keyword set can be written as:

$$\langle -\log P(w) \rangle = -\frac{1}{|K_i|} \sum_{w \in K_i} \log P(w) \tag{3}$$

While formally appropriate as a measure of novelty, the above quantity suffers from the disadvantage that its maximum attainable value, $\log(|M^i| + 1)$, is dependent on how many films have been released prior to the film under consideration. To yield a fair comparison between films irrespective of their position in the temporal order, we normalize the surprise associated with each keyword by the maximal attainable surprise ($\log(|M^i| + 1)$ for film i), and define the *elemental novelty* associated with film i as:

$$\mathcal{N}_E^i = -\frac{1}{|K_i|(-\log(|M^i|+1))} \sum_{w \in K_i} \log P(w) \tag{4}$$

Thus, $\mathcal{N}_E^i$ represents how close the average surprise for film i, as defined by Eq. 3, is to its maximum attainable value.

While Eq. 4 scores films based on the rarity or abundance of their individual plot-elements, it is agnostic to how rare or abundant the *combinations* of their plot-elements are. To capture the novelty associated with the combinations of keywords, we can define similarly to Eq. 4, the novelty resulting from the occurrence of specific keyword-pairs, triples and so on. Here we restrict our study of higher-order terms to keyword-pairs and formally write the *combinatorial* novelty for film i as:

$$\mathcal{N}_C^i = -\frac{1}{|K_i|(|K_i|+1)(-\log(|M^i|+1))} \sum_{u,v \in K_i} \log P(u, v) \tag{5}$$

where $P(u, v)$ is the probability of keywords u and v occurring together in a film in the set $M^i \cup \{i\}$ (defined similarly as for

individual keywords in Eq. 1). Both $\mathcal{N}_E^i$ and $\mathcal{N}_C^i$ have values in the range [0, 1], but capture distinct aspects of novelty generation. Thus observing trends in their evolution over time, not only gives us insights pertinent to specific events in the history of cinema, but also helps elucidate the degree to which elemental and combinatorial novelty contribute to the creation of new content.

Figure 4(a) shows the chronological evolution of elemental novelty over the period 1929–1998. To eliminate situations where a film with a small keyword set registers a very high (very low) novelty due to the rarity (abundance) of its few keywords, we only consider films with keyword sets of length greater than 10 (see SI Section 1.1). Films are chronologically ordered by the time of release, and the abscissa is simply the index i of the films, with the vertical dashed lines corresponding to the indices demarcating the beginning of a new decade. As stated earlier, novelty is bounded above by 1, and the median value of elemental novelty (shown in red) is well below this bound over the entire period. Some features in the evolution also bear pointing out. For example, an upward trend can be seen around the mid-1960s in both the yearly median, as well as the lower envelope of the time series, which agrees well with the documented birth of the American New Wave which brought with it a marked shift in themes, style and modes of production[24]. Interestingly, the period between 1929 and 1945, commonly referred to as the golden age of Hollywood, is not marked by an increase in or a stable value of

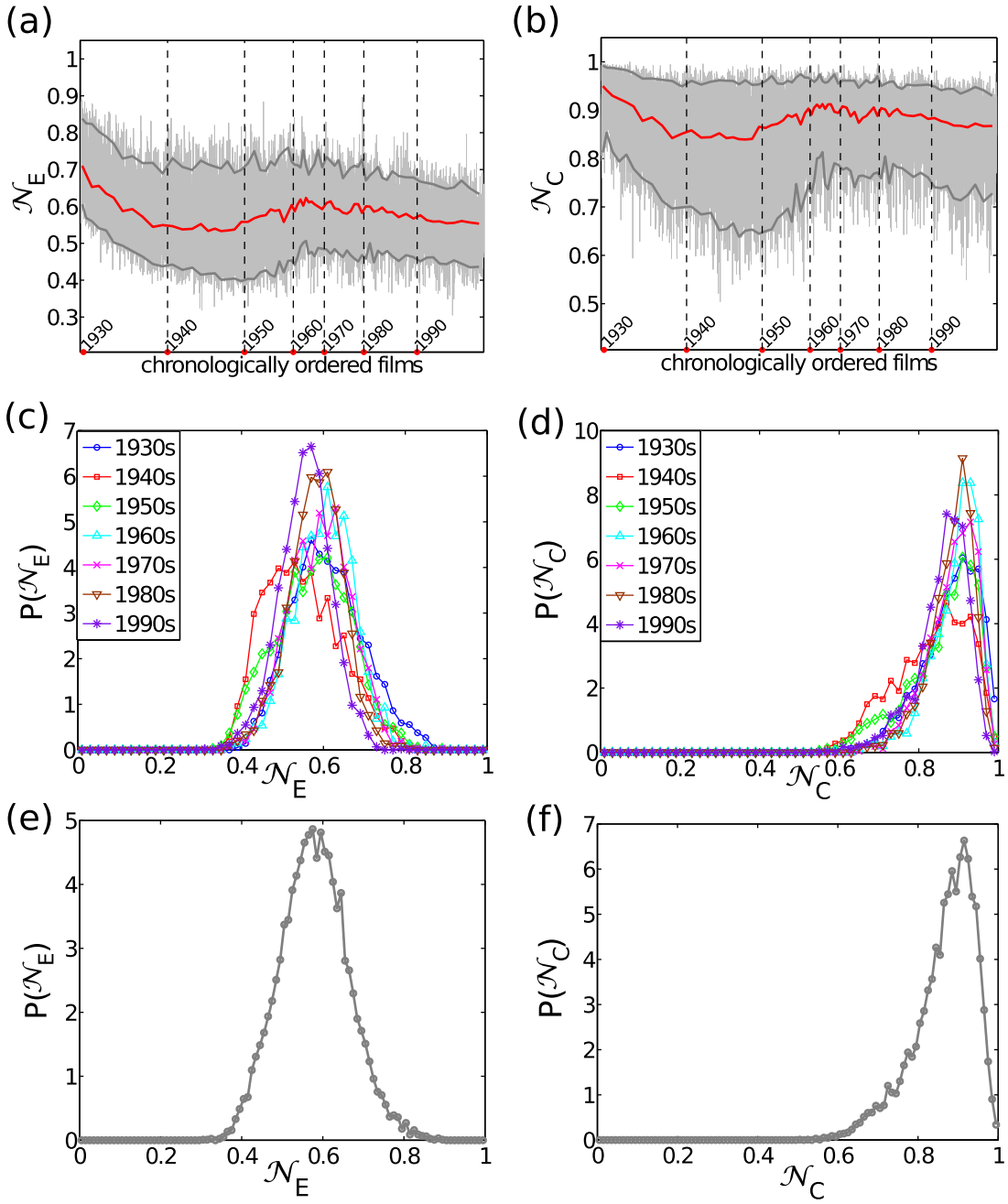

Figure 4 | The evolution of (a) elemental novelty and (b) combinatorial novelty for films between 1920 and 1998. The solid red curve shows the median yearly novelty, and the gray *envelope* curves show the novelty of the 5th and 95th percentile of films each year. Distributions of (c) elemental novelty and (d) combinatorial novelty by decade. Distribution of (e) elemental and (f) combinatorial novelty for the aggregated set of films released between 1929 and 1998.

median novelty, but rather by a subtle decline. This decline is likely a consequence of the practice of block booking prevalent in that period, which by virtually guaranteeing exhibition for any film as long as it came from a major studio, did little to de-incentivize the production of films with low novelty[2,24].

Figure 4(b) analogously shows the evolution of combinatorial novelty over the period, whose upper envelope in contrast to elemental novelty, consistently stays close to the maximum attainable value of 1. Gross features similar to those seen for elemental novelty can also be seen here; the median $\mathcal{N}_C$ rises in the 1960s and its variance decreases, while in contrast, the variance shows an increasing trend during the "golden age" between 1929 and 1945.

Figure 4(c) and (d) respectively show the probability density functions (pdf) of elemental-novelty and combinatorial-novelty for films in each of the 7 decades in the period considered. All the distributions are unimodal, differing slightly in their variances, but with their respective modes confined to the range between 0.53 and 0.63 for $\mathcal{N}_E$, and between 0.87 and 0.93 for $\mathcal{N}_C$. For each type of novelty, the similarity between individual decade-wise pdfs and the overall pdfs, Figs. 4(e), (f) respectively, hint at the possibility of some underlying novelty preferences governing which scripts are chosen for development into a feature film.

We also investigate the evolution of elemental and combinatorial novelty for films within specific genres, and these reveal trends unique to each of them. For example, Fig. 5(a), (b) show the evolution of novelties for films containing "Action" as one of their IMDb genre classes while Fig. 5(c) and (d) show the case for films under the "Sci-fi" genre. The median and the envelope curves of both $\mathcal{N}_E$ and $\mathcal{N}_C$ for the case of action films, show a sudden disruptive jump to higher values in the decade 1960–70. This is compatible with the thesis, based on studies by film historians, that elements comprising the modern action film genre originated with the James Bond franchise in the 1960s[33]. Similar plots for selected other genres are shown in Supplementary Figure 2.

Relationship between film novelty and revenue. Next, motivated by the Wundt-Berlyne curve, we investigate whether there is any relationship between the novelty of a film and the hedonic value derived from its consumption at an aggregate population level. We utilize the (inflation-adjusted) revenue generated by the film as a measure of its mass appeal (see Supplementary Text, Section 1.2), and measure a film's novelty only taking into account the films released in a 6 month window prior to its release (see Methods for details).

In Figure 6(a) we plot the mean revenue of films conditioned on elemental novelty (black circles). The overall shape of the resulting curve shows a resemblance to the Wundt curve (inset), with the mean revenue increasing systematically with novelty until a value of around 0.8, and declining thereafter. To get a better sense of the significance of this curve, we generate 50000 randomized versions of data where the values of revenues are shuffled. For every shuffled data set we obtain the mean revenue corresponding to each novelty bin, and then plot the 10^{th} and 90^{th} percentile of all mean revenues values obtained for each novelty bin (gray curves). A significant fraction of the true data points either straddle these curves, lie below the 10^{th} percentile curve, or lie above the 90^{th} percentile curve,

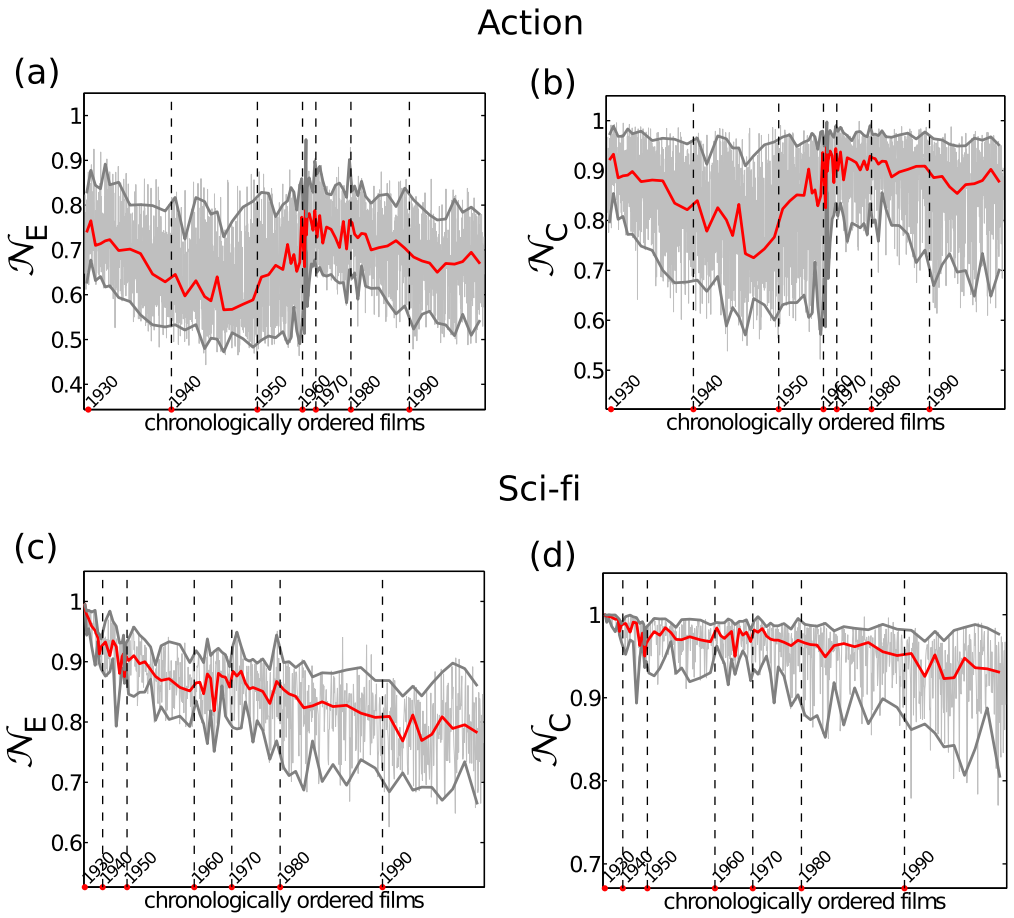

Figure 5 | (a) Elemental novelty and (b) combinatorial novelty for films containing 'Action' within their 'genre' field on IMDb. (c) Elemental novelty and (d) combinatorial novelty for films containing 'Sci-Fi' within their 'genre' field on IMDb. The solid red curve shows the median yearly novelty, and the gray envelope curves show the novelty of the 5th and 95th percentile of films each year.

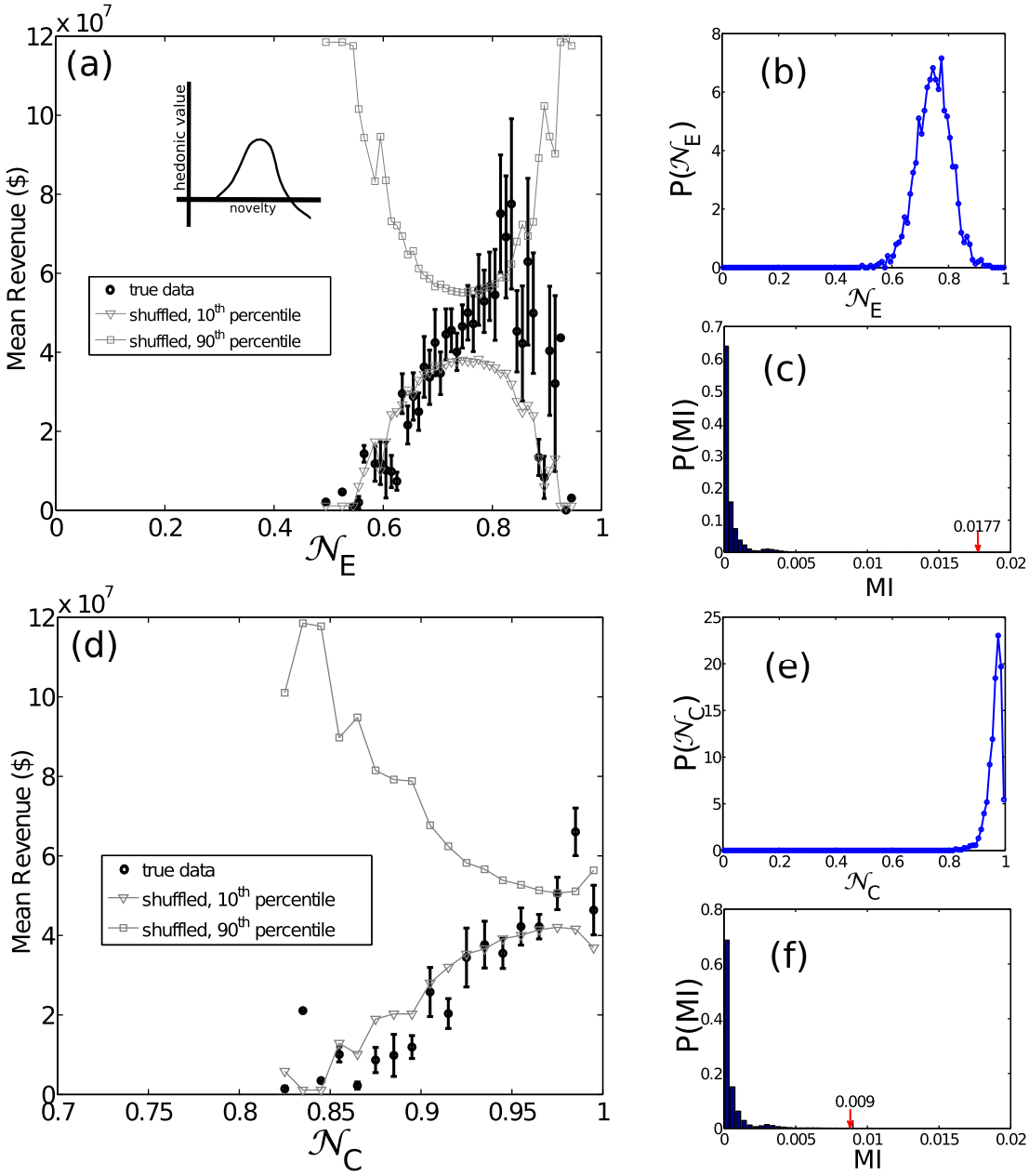

Figure 6 | (a) Mean inflation-adjusted revenue versus elemental novelty $\mathcal{N}_E$ is shown by the black circles, with vertical segments indicating standard-errors in the computed mean values. Also shown are the 10^{th} (gray triangles) and 90^{th} percentile (gray squares) of mean revenues obtained for 50000 randomized versions of the data. (b) The probability density function of $\mathcal{N}_E$ for the data used in (a). (c) The relative frequencies of mutual information values between $\mathcal{N}_E$ and mean revenue obtained for the randomized datasets, compared to the mutual information for the true dataset. (d) Mean inflation-adjusted revenue versus combinatorial novelty $\mathcal{N}_C$ (black curve) and the 10^{th} and 90^{th} percentile (gray triangles, gray squares respectively) of mean revenues obtained for 50000 randomized versions of the data. (e) The probability density function of $\mathcal{N}_C$ for the data used in (d). (f) The relative frequencies of mutual information values between $\mathcal{N}_C$ and mean revenue obtained for the randomized datasets, compared to the mutual information for the true dataset.

indicating that their respective probabilities of occurring purely due to chance is $\leq$10%. The declining portion of the curve is harder to conclusively argue for, due to a paucity of data points for the associated range of novelty, as evidenced by the probability density of novelty, Fig. 6(b). Irrespective of the precise nature of the relationship between novelty and hedonic value, we can investigate whether these two quantities exhibit a significant statistical dependance on one another. We do this by evaluating the Mutual Information (MI) (see Methods) between the two quantities, and comparing it to the values obtained from a permutation test. Specifically, we generate 50000 datasets where the revenue values are shuffled, and compute the MI between novelty and revenue for each shuffled dataset. Figure 6(c) shows that the MI for the true data (red arrow) is far from the tail of the distribution of MI values obtained using the shuffled datasets. More precisely, none of the shuffled datasets achieved a value equal to or greater than the true MI of 0.0177, indicating a p-value less than 2×10^{-5}.

Figure 6(d) shows the mean revenue of films as a function of their combinatorial novelty. Here the range of novelty values is much narrower (Fig. 6(e)), and the only discernible feature is a systematic increase in mean revenue as novelty increases. The MI between novelty and $\mathcal{N}_C$, as in the case of $\mathcal{N}_E$, is statistically significant as

indicated by a permutation test, with a p-value less than 2×10^{-5} (Fig. 6(f)).

Overall occurrence probabilities of keywords and keyword-pairs. Next, we study the probability distribution of plot-keywords over the entire set of films in the period between 1890 and 2011. Unlike the case for other corpora[3,26], the distribution does not follow Zipf's law as seen from the curvature present in the log-log plot of the cumulative probability distribution of usage frequency (Fig. 7(a)). Indeed, a stretched exponential fit obtained through maximum-likelihood-fitting[34,35] agrees well with the data (parameters provided in caption).

Any non-trivial process of plot generation would result in some keyword-pairs occurring more often than expected by chance, and others less often. To probe whether this is indeed borne out by the data, we compare the occurrence frequency of keyword pairs to the frequency obtained under the assumption that the constituent keywords are chosen independently of each other, in proportion to their respective occurrence probabilities. The results shown in Fig. 7(b) show a substantial difference between the true keyword-pair frequencies and those obtained under the independence assumption.

Finally, we present a visual depiction (Fig. 8) of the rise and fall of keywords that are associated with movies over the entire period from 1910 to 2011. Unlike a traditional time series plot (as in Fig. 3(a)) *streamgraphs* introduced in[36,37] provide a lucid graphical approach to simultaneously observing the growth and decline in the usage of different keywords (thickness of each "stream"), along with their relative usage in a given year (relative thickness of a stream in a cross section).

A prominently visible feature in Fig. 8(a) is the growth in the use of the keyword *independent-film* beyond 1955, presumably resulting from the demise of the studio system and marking the period when studios began forming partnerships with independent producers. Furthermore, until that time, the monopoly of the studios on the exhibition venues, strongly suppressed the visibility of independently produced films[24]. A notable feature in the action streamgraph (Fig. 8(c)) is the early dominance of the keyword *b-movie* and its decline in the 1950s. Indeed, between 1930 and 1950, action films mostly comprised of low-budget westerns created to fit the double feature programming format[38]. However, by the 1950s, with film audience numbers in decline as a result of the predominance of television, and with the end of the studio-system, the low-budget action film gradually declined in production and the genre as a whole underwent a redefinition in the 1960s[33].

Discussion

We have demonstrated that user-generated keywords coarsely characterizing a film, can provide a quantitative window into the evolution of novelty in films over a 70 year period. Specifically, the novelty scores defined here reveal both subtle trends in overall novelty evolution (Fig. 4) and disruptive changes in the evolution of specific genres (Fig. 5(a)). A notable feature of several evolution curves is an upward trend in novelty during the 1960s (Fig. 4(a), (b), and Fig. 5(a), (b)). Presumably, this corresponds to the widely held thesis[24] that the break-up of the studio system, the advent of competition from television, and the rise of several socio-political movements, all contributed in varying measures to the 1960s becoming a defining decade in the history of American cinema.

However, the fact that the overall distributions as well as the decade-wise distributions of $\mathcal{N}_E$ and $\mathcal{N}_C$ overlap significantly, suggests some strong constraints on the degree of novelty in films that eventually get made and released theatrically. This could be a manifestation of the inherent novelty preferences of the investors, or of risk-minimization based on some implicitly perceived inverted-U relationship between novelty and hedonic value. Indeed, the plot of $\mathcal{N}_E$ versus mean-revenue, Fig. 6(a) does lend some credence to the idea that the relationship between novelty and hedonic value resembles the Wundt-Berlyne behavior.

While this study has focussed on utilizing keywords to observe aggregate trends, there are several possible extensions that can be pursued in future work. The first is to attempt a refinement of novelty scores which takes into account the descriptive level of the keyword, an issue that is ignored in this study. For example, here we treat a keyword characterizing a high-level feature related to the production (for example *independent-film*) equivalently to a keyword which specifies a story-element (for example *murder*). A possible approach to alleviating this is by employing a probabilistic topic model like hierarchical latent Dirichlet allocation on the keyword set[39], and then defining a more finely resolved measure of novelty based on the obtained hierarchy of topics.

A second potential research direction is to analyze the utility of the novelty score discussed here or refinements of it to search and recommendation. Yet another application of such scores is in the area of artificial or computer-aided story generation[40] where ranking

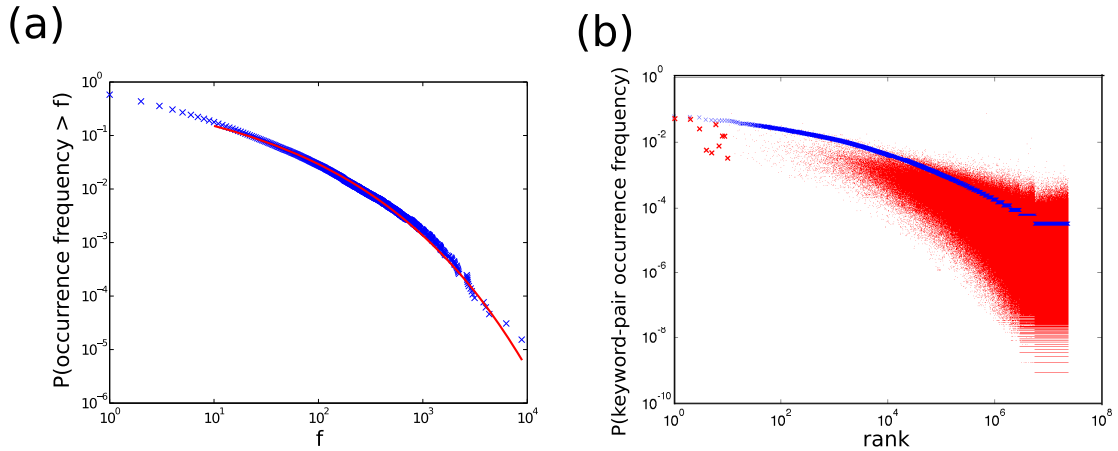

Figure 7 | (a) The cumulative probability, P(occurrence frequency $> f$) for keywords. The red line shows a fit corresponding to the cumulative probability for a stretched exponential distribution, $\exp(-\lambda f^{\beta})$ with parameters $\lambda = 1.0119$ and $\beta = 0.2716$. (b) The frequency of keyword-pair occurrence as a function of the rank of the pair (blue). For the keyword-pair corresponding to each rank, the probability of occurrence under the independence assumption is shown by a red dot. For the 10 highest ranked keyword-pairs, the probabilities of occurrence under the independence assumption are indicated by red crosses.

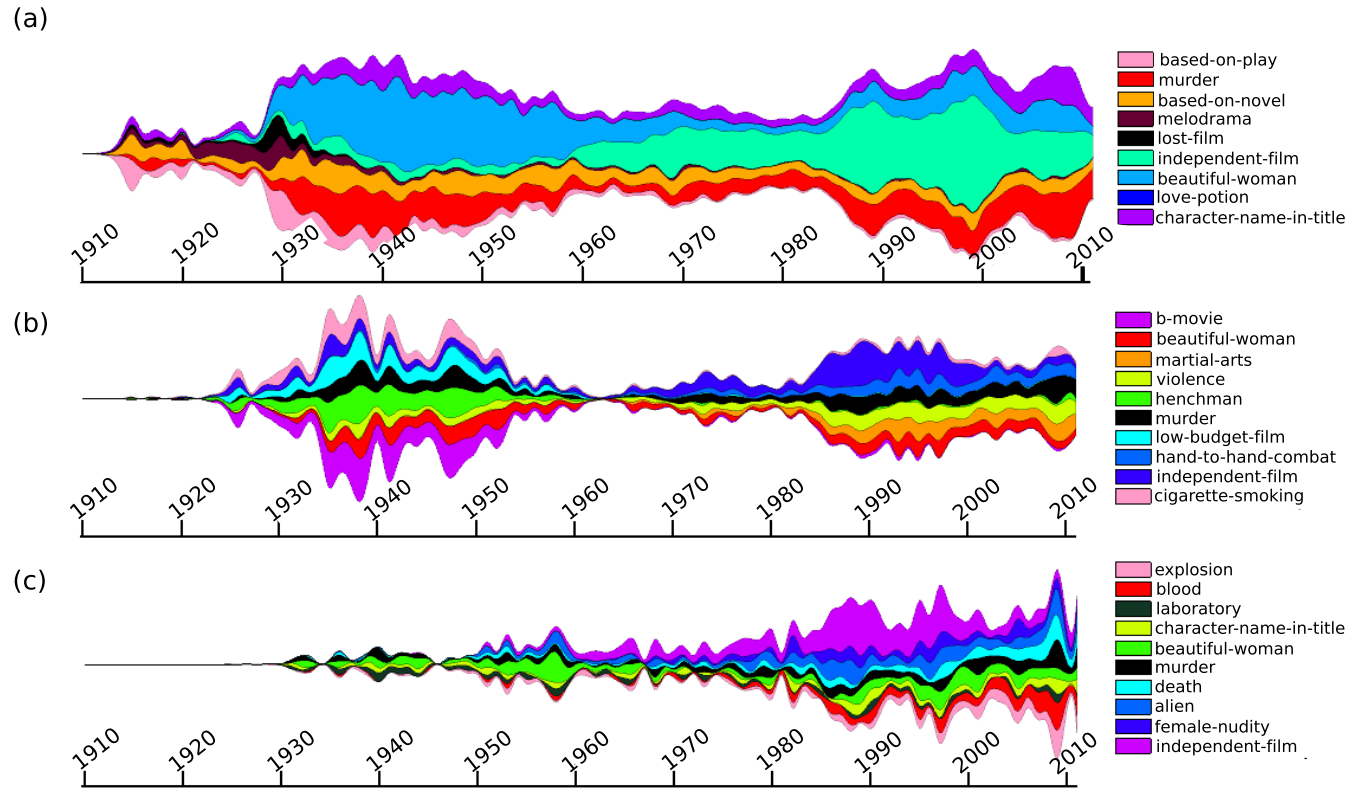

Figure 8 | Streamgraphs for most probable keywords occurring in (a) all films (b) action films and (c) science-fiction films. See Methods for details.

the novelty of plot-element combinations based on their prior probabilities could allow exploration in novel directions. Understanding aggregate novelty preferences may also provide insights into the viral spread and mass adoption (or lack thereof) of certain products and services, and is a research direction with valuable applications to marketing campaigns and social network based behavior-change initiatives. Furthermore, any venue offering the combined availability of crowdsourced data, the network between users providing tags, and their individual tagging behavior, provides the opportunity to segment the population on the basis of their novelty preferences, and design products and services tailored specifically to each segment.

Methods

Data collection and analysis. Data was obtained from IMDb (http://www.imdb.com/interfaces) as plain text data files in May 2012. Data was processed with Python scripts using the IMDbPY package (http://imdbpy.sourceforge.net/). First, all data items corresponding to films (not including straight-to-video releases, or TV movies) were extracted. Next, those items which had "Country" listed as 'USA' and "Language" listed as 'English' were extracted. Finally, all films with 'Adult', 'Short' or 'Documentary' under "Genre" were removed to leave us with the set under consideration. For more details, see Supplementary Text, Section 1.1.

Detrended fluctuation analysis. Detrended fluctuation analysis for a time series $y \equiv \{y_1, y_2, \cdots y_N\}$ involves the following steps:

(i) Mean-center the original time series: $\bar{y} \equiv \{y_1 - \langle y \rangle, y_2 - \langle y \rangle, \cdots, y_N - \langle y \rangle\}$ where

$$\langle y \rangle = \frac{\sum_{i=1}^{N} y_i}{N}$$

(ii) Generate a random walk z by summing up displacements corresponding to values in $\bar{y}$:

$$z_j = \sum_{i=1}^{j} \bar{y}_i$$

(iii) Partition the total number of steps in the walk (i.e., total number of elements in the original time series) into boxes of size L.

(iv) Within each box, compute the local trend $\bar{z}$ using a linear fit to the data. Compute the variance in the detrended fluctuations within each box and then compute the square root of its average over all boxes: $\sigma(L) \sim \sqrt{\left\langle (z(t) - \bar{z})^2 \right\rangle}$ (where the $\langle \cdots \rangle$ corresponds to an average over boxes, and the term within corresponds to the variance within a box.

(v) Repeat the process for different values of L and estimate the exponent α in the scaling $\sigma(L) \sim L^{\alpha}$.

Mutual information estimation. The mutual information between random variables x and y with marginal distributions $P(x)$ and $P(y)$ respectively and joint distribution $P(x, y)$ is defined as:

$$I = \sum_{x,y} P(x,y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right)$$

In the absence of a knowledge of a specific form for the relationship between variables, mutual information is a useful signifier of the presence or absence of dependencies between variables x and y[41]. The estimation of mutual information between two continuous variables with a finite number of observations is a well-studied problem. We utilize a method proposed in[42,43] and an implementation of the same provided by Zbynek Koldovsky.

Novelty and hedonic value. The following pertains to the data and methods used for Fig. 6. Budgets and revenues generated from theatrical exhibition are present for 1680 films in the period under consideration. We adjust for inflation all dollar amounts that have a reporting year associated with them based on the cumulative price index table for the year 2011. To strike a balance between having a sufficiently large number of films to analyze, and minimizing the disparities in the exhibition capabilities of films considered, we restrict our analysis to films with a inflation adjusted budget of at least 1 million dollars (see SI, Section 1.2 for further details). Finally, to account for the fact that novelty as perceived by a general audience largely involves comparison to films released over a short period in the past (rather than the over the entire duration that cinema has been around), we compute $\mathcal{N}_E$ and $\mathcal{N}_C$ for a film i, only considering films which were released in the 6 months preceding the month of its release.

Streamgraphs. A "stream" for a keyword was generated using the number of occurrences of the keyword for each year in the period. The resulting signal was smoothed using spline interpolation. A stacked graph was generated and to guarantee symmetry about the Y axis, the baseline was displaced in proportion to the total width of the stack as described in[36,37]. For Fig. 8(a) we use the set of

keywords obtained from the union of the most frequently used keyword for each year in the period. This set contains 9 unique keywords. For streamgraphs shown in Figs. 8(b) and (c) for films belonging to the action and science-fiction genres respectively, keyword sets were chosen using a similar procedure as for Fig. 8(a) but were additionally pruned to retain only the 10 keywords with the highest average usage-frequency over the period.

1. 2011 Theatrical Market Statistics, Motion Picture Association of America http://www.mpaa.org/policy/industry (Accessed March 15, 2013).
2. DeVany, A. *Hollywood Economics: How Extreme Uncertainty Shapes the Film Industry* (Routledge, 2003).
3. Michel, J.-B. *et al*. Quantitative analysis of culture using millions of digitized books. *Science* **331**, 176–182 (2011).
4. Halpin, H., Robu, V. & Shepherd, H. The complex dynamics of collaborative tagging. *Proceedings of the 16th international conference on World Wide Web*, WWW '07, 211–220 (ACM, New York, NY, USA, 2007).
5. Cattuto, C., Loreto, V. & Pietronero, L. Semiotic dynamics and collaborative tagging. *Proc. Natl. Acad. Sci.* **104**, 1461–1464 (2007).
6. Brooks, C. H. & Montanez, N. An analysis of the effectiveness of tagging in blogs. *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs '06*, 9–14 (2006).
7. McAuley, J. & Leskovec, J. Image labeling on a network: using social-network metadata for image classification. *Proceedings of the 12th European conference on Computer Vision - Volume Part IV*, ECCV'12, 828–841 (Springer-Verlag, Berlin, Heidelberg, 2012).
8. Farooq, U. *et al*. Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. *GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work*, 351–360 (ACM, New York, NY, USA, 2007).
9. Levy, M. & Sandler, M. A semantic space for music derived from social tags. *ISMIR'07*, 411–416 (2007).
10. Shepitsen, A., Gemmell, J., Mobasher, B. & Burke, R. Personalized recommendation in social tagging systems using hierarchical clustering. *Proceedings of the 2008 ACM conference on Recommender systems*, RecSys '08, 259–266 (ACM, New York, NY, USA, 2008).
11. Szomszor, M. *et al*. Folksonomies, the Semantic Web, and Movie Recommendation. *Bridging the Gap between Semantic Web and Web 2.0 (SemNet 2007)* 71–84 (2007).
12. Gupta, M., Li, R., Yin, Z. & Han, J. Survey on social tagging techniques. *SIGKDD Explor. Newsl.* **12**, 58–72 (2010).
13. Milicevic, A. K., Nanopoulos, A. & Ivanovic, M. Social tagging in recommender systems: a survey of the state-of-the-art and possible extensions. *Artif. Intell. Rev.* **33**, 187–209 (2010).
14. Mestyán, M., Yasseri, T. & Kertész, J. Early prediction of box-office success based on Wikipedia activity big data. *PLoS ONE* **8**(8): e71226 (2013).
15. Koestler, A. *The Act of Creation* (Penguin Books, 1990).
16. Boden, M. A. *The Creative Mind: Myths and Mechanisms* (Routledge, 2003).
17. Hoftsadter, D. R. *Metamagical Themas*, chap. 12 (Basic Books, 1985).
18. Sluckin, W., Colman, A. M. & Hargreaves, D. J. Liking words as a function of the experienced frequency of their occurrence. *Brit. J. Psychol.* **71**, 163–169 (1980).
19. Berns, G. *Satisfaction: The Science of Finding True Fulfillment* (Henry Holt and Co., 2005).
20. Anand, P. & Holbrook, M. B. Chasing the Wundt curve: an adventure in consumer esthetics. *Adv. Consum. Res.* **13**, 655–657 (1986).
21. Saunders, R. & Gero, J. S. How to study artificial creativity. *Proceedings of the 4th conference on Creativity & cognition*, C&C '02, 80–87 (ACM, New York, NY, USA, 2002).
22. Berlyne, D. E. *Aesthetics and Psychobiology* (Appleton-Century-Crofts, 1971).
23. Berlyne, D. Novelty, complexity, and hedonic value. *Percept. Psychophys.* **8**, 279–286 (1970).
24. Ellis, J. C. & Wexman, V. W. *A History of Film* (Allyn and Bacon, 2002).
25. Mcdonald, P. & Wasko, J. (ed.) *The Contemporary Hollywood Film Industry* (Wiley-Blackwell, 2008).
26. Petersen, A. M., Tenenbaum, J., Havlin, S. & Stanley, H. E. Statistical laws governing fluctuations in word use from word birth to word death. *Sci. Rep.* **2**, 313; doi:10.1038/srep00313 (2012).
27. Collins, R. M. *Transforming America: Politics and Culture During the Reagan Years* (Columbia University Press, 2009).
28. McKeage, K. K. Materialism and self-indulgences: Themes of materialism in self-gift giving. Rudmin, F. W. & Richins, M. (eds.) *Meaning, Measure, and Morality of Materialism, Provo, UT: Association for Consumer Research*, 140–146 (1992).
29. Peng, C. K. *et al*. Mosaic organization of DNA nucleotides. *Phys. Rev. E* **49**, 1685+ (1994).
30. Sedgwick, J. & Pokorny, M. The film business in the United States and Britain during the 1930s. *Econ. Hist. Rev.* **58**, 79 (2005).
31. Jewell, R. *The Golden Age of Cinema: Hollywood*, 1929–1945 (Wiley-Blackwell, 2007).
32. Cover, T. M. & Thomas, J. A. *Elements of information theory* (Wiley-Interscience, 2006).
33. Chapman, J. *Licence to Thrill: A Cultural History of the James Bond Films (Cinema and Society)* (I. B. Tauris, 2008).
34. Virkar, Y. & Clauset, A. Power law distributions in binned empirical data. arXiv:1208.3524.
35. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009).
36. Havre, S., Hetzler, B. & Nowell, L. Themeriver: Visualizing theme changes over time. *Proceedings of the IEEE Symposium on Information Visualization 2000*, INFOVIS '00, 115 (IEEE Computer Society, Washington, DC, USA, 2000).
37. Byron, L. & Wattenberg, M. Stacked graphs - geometry & aesthetics. *IEEE T. Vis. Comput. Gr.* **14**, 1245–1252 (2008).
38. Nachbar, J. *Focus on the Western* (Prentice-Hall, 1974).
39. Blei, D. M., Griffiths, T. L., Jordan, M. I. & Tenenbaum, J. B. Hierarchical topic models and the nested chinese restaurant process. *Advances in Neural Information Processing Systems*, 2003 (MIT Press, 2004).
40. Veale, T., Gervás, P., Pérez, Y. & Pérez, R. Computational creativity: A continuing journey. *Minds Mach.* **20**, 483–487 (2010).
41. Steuer, R., Kurths, J., Daub, C. O., Weise, J. & Selbig, J. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* **18**, S231–S240 (2002).
42. Darbellay, G. A. & Vajda, I. Estimation of the information by an adaptive partitioning of the observation space. *IEEE T. Inform. Theory* 1315–1321 (1999).
43. Darbellay, G. A. & Tichavsky, P. Independent component analysis through direct estimation of the mutual information. *ICA'2000 Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, Helsinki, Finland*, 69–75 (2000).

Acknowledgments

This work was supported in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 and by the Office of Naval Research Grant No. N00014-09-1-0607. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies either expressed or implied of the Army Research Laboratory or the U.S. Government. S.S. thanks B.K. Szymanski, G. Korniss and A. Asztalos for critical readings of the manuscript, and valuable comments and suggestions. S.S. thanks Y. Virkar for a Matlab implementation of the maximum-likelihood-fit of a stretched exponential function to binned data, and Zbynek Koldovsky and Petr Tichavsky for a Matlab implementation of the estimation of mutual information using adaptive partitioning.

Author contributions

S.S. designed the research, performed the analysis of data and wrote the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The author declares no competing financial interests.

How to cite this article: Sreenivasan, S. Quantitative analysis of the evolution of novelty in cinema through crowdsourced keywords. *Sci. Rep.* **3**, 2758; DOI:10.1038/srep02758 (2013).

SCIENTIFIC REPORTS

DOI: 10.1038/srep03972

SUBJECT AREAS:
APPLIED PHYSICS
COMPUTATIONAL SCIENCE
SCIENTIFIC DATA
APPLIED MATHEMATICS

SCIENTIFIC REPORTS:
3 : 2758
DOI: 10.1038/srep02758
(2013)

Published:
26 September 2013

Updated:
29 January 2014

CORRIGENDUM: Quantitative analysis of the evolution of novelty in cinema through crowdsourced keywords

Sameet Sreenivasan

This Article contains typographical errors in Equations (4) and (5).

Equation (4) should read:

$$\mathcal{N}_E^i = -\frac{1}{|K_i|(\log(|M^i|+1))}\sum_{w\in K_i}\log P(w)$$

Equation (5) should read:

$$\mathcal{N}_C^i = -\frac{1}{|K_i|(|K_i|-1)(\log(|M^i|+1))}\sum_{u,v\in K_i}\log P(u,v)$$