

Quantifying and predicting success in show business

Oliver E. Williams,¹ Lucas Lacasa,¹ Vito Latora^{1,2,3,4}

¹*School of Mathematical Sciences, Queen Mary University of London, London, E1 4NS, United Kingdom.*

²*The Alan Turing Institute, The British Library, London NW1 2DB, UK,*

³*Dipartimento di Fisica ed Astronomia, Università di Catania and INFN, I-95123 Catania, Italy.*

⁴*Complexity Science Hub Vienna (CSHV), Vienna, Austria.*

*

Abstract. *In certain artistic endeavours –such as acting in films and TV, where unemployment rates hover at around 90%– sustained productivity (simply making a living) is probably a better proxy for quantifying success than high impact. Drawing on a world-wide database, here we study the temporal profiles of activity of actors and actresses. We show that the dynamics of job assignment is well described by a ‘rich-get-richer’ mechanism and we find that, while the percentage of a career spent active is unpredictable, such activity is clustered. Moreover, productivity tends to be higher towards the beginning of a career and there are signals preceding the most productive year. Accordingly, we propose a machine learning method which predicts with 85% accuracy whether this “*annus mirabilis*” has passed, or if better days are still to come. We analyse actors and actresses separately, also providing compelling evidence of gender bias in show business.*

*E-mails: o.e.williams@qmul.ac.uk — l.lacasa@qmul.ac.uk — v.latora@qmul.ac.uk

1 Introduction

“*It’s feast or famine in showbiz.*” - *Joan Rivers*. A sentiment likely to be echoed by many would-be stars of the silver screen. But for those that feast the rewards are, at least thought to be, worth the risk. The so-called science of success has recently uncovered many features of the careers of academics (1), artists (2), and all manner of other individuals whose output can be effectively assessed over the course of their working life (3–5). For instance, in the world of scientific research it has revealed the unpredictability of the location of an academics most impactful work (1), showing that even such prestigious awards as Nobel prizes, which usually occur later in a career (6), are underpinned by research papers that are located randomly and uniformly throughout the ordered list of papers in the career of the awardee. On the other hand, the anatomy of funding and collaborations in universities has revealed “rich clubs” of leading institutions, and suggested that such patterns of collaborations contribute greatly to the success of these institutions, as measured in terms of over-attraction of available resources and of breadth and depth of their research products (7). Studies of innovation in industry across different countries have found that the commercial success of manufacturing plants is far more closely related to intra-group links than external ties (8). Strikingly, these features can be common across multiple areas; the Matthew effect (9, 10), or the rich-get-richer phenomenon, and the recently discovered presence of “hot streaks” (11), are not restricted to isolated cases. With regards to success, a great deal of work has been done in assessing impact (1, 12), the distribution of standout or landmark works (13, 14), whether these are related to the age of the individual in question (15, 16), how impact can be assessed in the long term (17), and even prediction of future successes (18, 19). Indeed the fortunes of both films and the actors and actresses that make them have been studied in some specific ways (17, 20–22). These studies do not however address the question that interests those who are not already on the higher rungs of the ladder of success: how can one avoid the famine and build a sustainable career in acting?

The aim of this work is to use a data-driven approach in order to define, quantify and even predict the success of actors and actresses in terms of their ability to maintain a steady flow of jobs. Drawing on the International Movie Database (IMDb), an online database of information related to films, television programs and home videos (23), we study the careers of millions of actors from several countries worldwide, from the birth of film in 1888 up to the present day. Each career is viewed as a profile sequence: the yearly time series of acting jobs in films or TV series over the entire working life of the actor or actress (this is similar in spirit to the approach used in (24) to explore scientific productivity). Note that all acting jobs are considered, regardless of salary, role, screen time, or the impact of the work. The statistical analysis of such a large number of profile sequences allows us to derive some general properties of the actors activity patterns. In particular, we look at several quantities of interest such as *career length*, *productivity* (defined as the number of credit jobs in a year or in the entire career of an actor) and position of the *annus mirabilis*, defined as the year with the largest number of credited jobs. We also explore possible emergence of gender inequality in these properties.

The first message that emerges from our quantitative analysis is that one-hit wonders, i.e. actors whose career spans only a single year, are the norm rather than the exception. Long career lengths and high activity are found to be exponentially rare, suggesting a scarcity of resources in the acting world. These results are in agreement with previously collected evidence, pointing to the fact that

unemployment rates in actors hover around 90%, and that as low as 2% of actors are able to make a living out of acting (25). We also observe that that this dramatic scarcity unequally applies to actors and actresses, providing compelling evidence of gender bias. Moreover, the total productivity of an actor’s career is found to be power-law distributed, with most actors having very few jobs, while a few of them have more than a hundred. This indicates a rich-get-richer mechanism underpinning the dynamics of job assignments, with already scarce resources being allocated in a heterogeneous way. All of this suggests that, while activity and sustained productivity are by definition measures of performance (26), they should in this context be considered as a proxy for success. Only a select few will ever be awarded an Oscar, or have their hands on the walk of fame, but this is not important to the majority of actors and actresses who simply want to make a living. It is the continued ability to work (as opposed to prestige) that is most likely to ensure a stable career. For these reasons we propose that predictions of success in show business should be focused on activity and productivity. Observe at this point that performance is usually conflated with success (26). While performance is objectively measured in terms of an individual’s actions, and is typically bounded, success is traditionally measured by recognition, i.e. in terms of impact, and is a collective phenomenon which is unbounded. Notwithstanding, the severe scarcity of resources in show business forces us to redefine an actor’s success, not in terms of popularity or impact, but in terms of activity and productivity as discussed above. Incidentally, note also that being credited on IMDb is to a certain extent funnelled by recognition mechanisms such as popularity – a producer might offer the job to the actor who had the best audition or to the one who has more followers on Instagram–, so productivity is not only, strictly speaking, a performance-driven indicator.

Motivated by these results, we then address the questions that interest the majority of working actors and actresses. Questions such as “am I going to get another paid job?” or “is this year going to be my best?”. We first show that efficiency, defined as the ratio between the total number of active years and the career length, is unpredictable, as there is no evident correlation between these two things. This is in line with recent studies (1) pointing out that the most impactful pieces of work in scientific disciplines are equally likely to be located in any position throughout the entirety of an individuals output, and is therefore not predictable. Nevertheless, we here, surprisingly, find distinctive features in their temporal arrangement. In particular, we find that actor careers are clustered in periods of high activity (hot streaks) (11) combined with periods of latency (cold streaks). Moreover, we discover that the most productive year (annus mirabilis) for both actors and actresses is located towards the beginning of their career, and that there are clear signals preceding and following the location of the annus mirabilis of an individual. Altogether, these unexpected results lead us to conclude that prediction is possible in theory. Finally, we validate this hypothesis by building a statistical learning model which predicts the location of the most productive year, finding that we can, with up to 85% accuracy, tell whether an actor’s career has reached its most productive year yet or not.

2 Results

We study the careers of 1, 512, 472 actors and 896, 029 actresses as recorded on IMDb as of January 16th, 2016, including careers stretching back to the first recorded movie in 1888. The career of each actor a is characterised by his/her track record, which consists of a set of pairs of numbers representing respectively each year when actor a was credited in IMDb, and the number of different credits in that

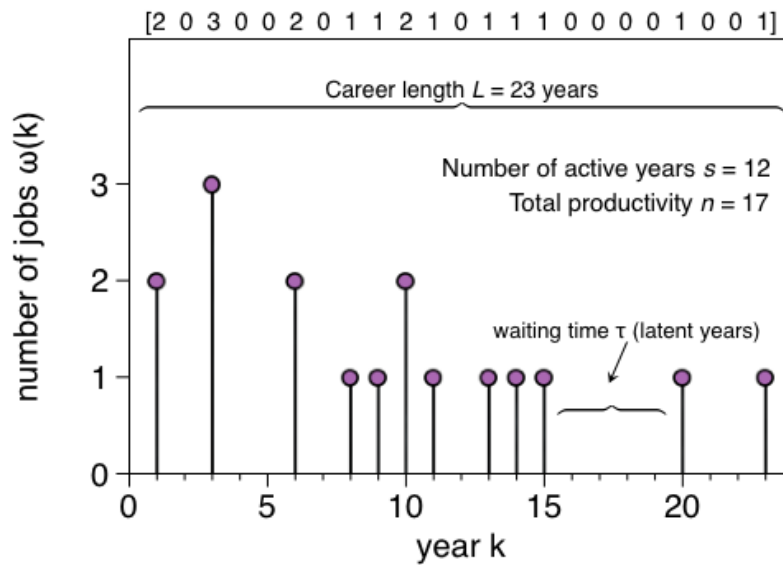


Figure 1: **Career activity pattern of an actor.** The yearly productivity of a given actor, measured as the total number of IMDb credited jobs in each year, is reported from the first to the last year of the actor activity. Shown is the case of an actor whose career spanned $L = 23$ years and who was credited a cumulated $n = 17$ different jobs in $s = 12$ years. From the yearly productivity we can construct the actor *profile sequence* w_k , with $k = 1, \dots, L$, shown in brackets above the plot, which can be modelled as a stochastic marked point process.

year. As credits we count the number of acting jobs in films and/or TV series. A sketch of the typical activity pattern of an actor is reported in Figure 1, showing the yearly credits from the first to the last year of thir career. Notice that there are not only active years, where the actor has credited jobs in IMDb, but also latent years with no recorded jobs. We therefore fill the latent years with zeros and construct the profile sequence $\{w_k\}_{k=1}^L$ of each actor a as depicted in the top part of Figure 1. The quantity w_k denotes the actor’s local productivity in year k , i.e. the number of credited jobs in that year. The length of an actor’s career is defined as the number of years between the first and the last active year (inclusive), and is denoted as L . The total number of active years s is from now on referred to as the activity of an actor. Since a career can have latent years intertwined with active ones we must have $s \leq L$, moreover $L - s$ is the number of latent years. By definition we have: (i) $L \geq 1$, (ii) $s \geq 1$ and (iii) $s = 1 \Leftrightarrow L = 1$.

Finally, we define the total productivity n of an actor, as the cumulated number of credited jobs, $n = \sum_{k=1}^L w_k$. The annus mirabilis (AM) of a given actor is defined as the year where the actor was credited with the largest number of works in IMDb: $AM = m$, where m is such that $w_m = \max\{w_k\}_{k=1}^L$. In the case that this m is not unique we take the final such year: $AM = \max\{m\}$.

2.1 Career lengths and one-hit wonders

We start our analysis by exploring the statistics of the career length L . In panel (a) of Figure 2 we plot in a semi-log scale the empirical distribution of career lengths $P(L)$, for both actors and actresses finding that the tail is well fitted by an exponential distribution. By construction, $P(L = 1) = P(s = 1)$ and this quantity represent the percentage of one-hit wonders i.e. of actors whose career started and ended, according to IMDb, in the same year. Interestingly, we find that the percentage of such cases is extremely high (around 69% for males and 68% for females) and deviates from the otherwise decaying exponential distribution. This sharp deviation highlights that one-hit wonders are not an exception in show business, but, on the contrary, are the norm (37). A zoom of the distribution in the range $L \in [2, 10]$ is reported in the inset of (a), revealing systematic differences between actors and actresses, suggesting that it is consistently more common to find (non-one-hit wonder) actresses with shorter career lengths than actors. We have indeed performed a model selection experiment which confirms that gender bias is statistically significant (see Supplementary Note 1 for details).

The empirical probability distribution of activities, displaying the probability of sampling an actor that worked in s years, is shown in panel (b) of Figure 2 in a semi-log scale. Most of the actors and actresses are only active in a single year ($s = 1$), as by default $s = 1 \mapsto L = 1$. The probability of finding actors with large activity, i.e. those that have worked in many different years, decays exponentially fast. This exponential decay mimics the similar decay in the probability of finding long career lengths and altogether are the basis for claiming a scarcity of resources in show business, i.e. there are many more actors/actresses than job offers (27). This lack of resources naturally leads to a question: how are they allocated? We address this question in the next section.

2.2 Productivity and the rich-get-richer phenomenon

The right panel of Figure 2 shows the empirical distributions of total productivity $P(n)$, reporting the normalized numbers of actors or actresses with n appearances in movies or TV series over their

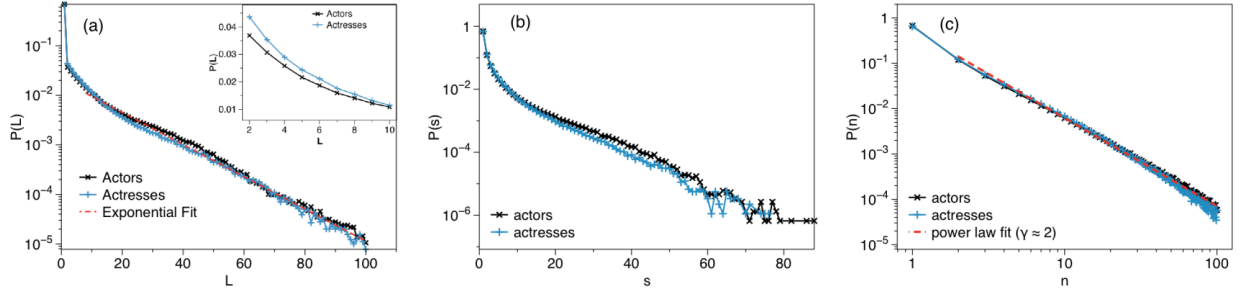


Figure 2: Career length, activity and productivity distributions. (a) The probability $P(L)$ that an actor or an actress has a career of length L , estimated by computing the frequency histogram of the number of years between the first and the last recorded entry on IMDb. $P(1)$ measures the abundance of “one-hit wonders”, namely the actors or actresses with IMDb records in a single year. A zoom for $L \in [2, 10]$ in the inset shows that careers extending between 2 and 10 years are proportionally more frequent in women than in men. (b) Activity distribution $P(s)$ estimated by computing the frequency histogram of the number of working years within each career ($s \leq L$). Curves for actors and actresses are very similar and both exhibit a clear exponential tail, implying a ‘scarcity of resources’. (c) Log-log plot of the total productivity distributions $P(n)$ for actors (black) and actresses (blue). Both curves decay as a power law $P(n) \sim n^{-\gamma}$, where $\gamma \approx 2$, revealing a Zipf’s law for the total number of acting jobs.

careers. While the career length distribution $P(L)$ and the activity distributions $P(s)$ are well fitted in their tails by an exponential law, the function $P(n)$ decays more slowly and can be fitted by a power law $P(n) \sim n^{-\gamma}$ with exponent $\gamma \approx 2$. Notice that similar behaviours have already been found in the context of two-mode actor-movie networks and of other systems that can be modelled as bipartite graphs (28). A power law in the distribution of total productivity implies also the existence of scaling in the rank-frequency distribution of productivity. It is indeed well known that observing a power-law distribution with exponent γ for the abundance of some variable is equivalent to obtaining a power-law scaling for the frequency of the variable that appears with rank r : $f(r) \sim r^{-\alpha}$ (29). The exponents of the two scaling laws are mathematically related via $\alpha = 1/(\gamma - 1)$. The celebrated Zipf’s law refers to the particular case of an exponent $\alpha \approx 1$, which is indeed the case here. In turn, the emergence of a Zipf’s law for the rank-frequency distribution of the total productivity of an actor suggests a possible mechanistic explanation for our observations. Many different proposals for the mechanism underpinning the emergence of a Zipf’s law, and several names for the phenomenon itself, have been put forward in various contexts, including the Simon-Yule process, the mechanism of preferential attachment, the Matthew effect, the Gibrat principle, rich get richer, etc.

In this context, we can suggest a possible mechanism for the onset of a power-law distribution for the total productivity in terms of a rich-get-richer phenomenon. Let us consider a generative model of a bipartite graph whose two sets of nodes represent respectively actors and movies. Actors acquire new links to movies, thus increasing their productivity, if they get a role in those movies. Suppose all actor nodes start with zero edges and acquire their first edges only according to a fitness, that is initially assigned at random or on some hypothetical intrinsic acting skill. When more movie nodes enter the network, actor nodes that acquire new edges gain popularity and this, in turn, increases

their fitness. As it is well known that producers are more keen to offer a role to popular actors, actor nodes with high fitness are more likely to attract new edges. This leads to a multiplicative effect which clearly expresses the rich-get-richer phenomenon; actors with many job assignments will have a higher chance of working even more than actors with low productivity. In conclusion, the same rich-get-richer mechanism, which is at the heart of networks with power-law degree distributions (30–33), can also be the cause of the observed power laws in the total productivity of movie actors. This result is not at all unexpected, after all, the more well-known an actor is, the more likely producers will want him or her in their next film, if only for commercial purposes. What is perhaps dramatic about this observation is that it is well known that rich-get-richer effects are rather arbitrary and unpredictable, as large hubs can evolve out of unpredictable and random initial fluctuations which have been amplified, and not based on any particular intrinsic fitness (33) (such as acting skills). Quoting Easley and Kleinberg: “*if we could roll time back 15 years, and then run history forward again, would the Harry Potter books again sell hundreds of millions of copies, or would they languish in obscurity while some other works of children’s fiction achieved major success?*”. As a matter of fact, it seems likely that across different parallel universes productivity would still have a power-law distribution, but it is far from clear that the most productive actors would always be the same. Interestingly, this hypothesis has recently been validated in an online social experiment for the case of musical popularity (34). In summary, productivity is probably the variable every actor aims to maximise, but these results suggest that boosting productivity can be more of a network effect (35, 36) than a consequence of acting skills.

2.3 Efficiency is unpredictable

In Figure 2 we observed that career length L and activity s are variables which are both exponentially distributed, indicating a scarcity of resources. In this section we further explore whether the two quantities L and s are correlated. We first define an actor’s efficiency as the ratio s/L of active years over the entire career, and we investigate how the efficiency is distributed. The results reported in Supplementary Figure 1, show that: (1) the efficiency distribution drops rapidly as s/L approaches either zero or one –i.e. most actors and actresses have intermediate values of efficiency– and that (2) for middle-range efficiency the distribution is essentially uniform (see Supplementary Note 2 for additional details). This suggests that efficiency is not predictable and that, for middle-range efficiency, the only correlations that emerge between the activity s and the career length L come from the fact that, by construction, $s \leq L$. To further validate this, we performed a scatter plot of s versus L for all actors and actresses, and computed the Pearson correlation coefficient, then compared this to the correlation coefficient of a null model generated by randomly extracting values of L and s from the pool of career profiles, ensuring that $L \geq s$ (Supplementary Figure 2). For actors, s and L exhibit a Pearson correlation coefficient $r \approx 0.69$, whereas in the null model we obtained $r_{\text{null}} \approx 0.6$. In the case of actresses we found $r \approx 0.69$ and $r_{\text{null}} \approx 0.58$. As expected, s and L are indeed correlated quantities, but the correlations can almost entirely be explained by a null model. In other words, for intermediate ranges there are no additional correlations between length and activity: the activity of actors cannot therefore be predicted by their career length, and we can conclude that the efficiency is an unpredictable quantity.

2.4 Actors careers are clustered in hot and cold streaks

To understand the temporal arrangement of active years within the profile sequence of a given actor, we now consider the statistics of waiting times. A waiting time τ is defined as the time elapsed (in years) between two active years (equivalently, a waiting time is a collection of successive latent years), and its statistics provide a classical way to analyse the presence of memory and bursts in time series (38, 39). We have estimated the waiting time distribution $P(\tau)$ for actors and actresses, discarding those with short career lengths, $L < 10$ years, to avoid a lack of statistics. To estimate this distribution, for each actor (actress) we count how frequently one observes waiting times of a certain duration τ , and normalize the accumulated frequencies. This process will inevitably introduce finite size biases since, for short career lengths, we are more likely to find short waiting times, simply because there is no room for long ones. For a proper comparison we therefore have also computed the distribution for a randomized null model $P_{\text{null}}(\tau)$ where all of the profile sequences have been shuffled (while keeping the first event w_1 and the last event w_L unaltered). A lack of temporal correlations would imply $P_{\text{null}}(\tau) = P(\tau)$, whereas systematic differences suggest the onset of temporal correlations in the activity of actors. In panel (a) of Figure 3 we report the difference $P(\tau) - P_{\text{null}}(\tau)$ as a function of τ .

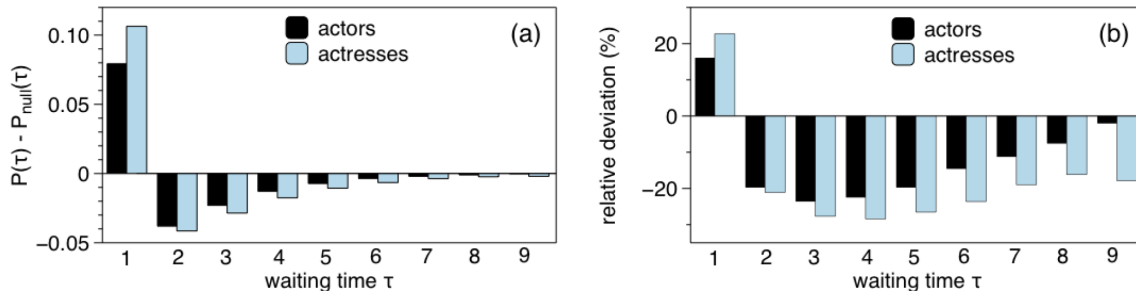


Figure 3: **Waiting time distribution.** (a) Difference $P(\tau) - P_{\text{null}}(\tau)$ between the waiting time distribution in the profile sequences and in a randomised null model, for actors (black bars) and actresses (blue bars). Systematically short waiting times, $\tau = 1$, are overrepresented with respect to the null model, while the opposite is true for intermediate waiting times $\tau > 1$. (b) The percentage relative difference $[P(\tau) - P_{\text{null}}(\tau)] \cdot 100/P_{\text{null}}(\tau)$ reveals a notable difference between actors and actresses: cold streaks fade away faster for actors.

For both actors and actresses, we systematically find $P_{\text{null}}(\tau = 1) < P(\tau = 1)$, and $P_{\text{null}}(\tau > 1) > P(\tau > 1)$, that is, active years are more clustered than they would be by chance, and hence the same is true of periods of inactivity. This means that the profile sequence shows *clustering* and is composed of bursts of activity (hot streaks) where actors and actresses are more likely, than would be expected by chance, to work in a year if they worked the year before ($\tau = 1$). This result is in agreement with recent findings in other creative jobs in science and art (11). Additionally, these hot streaks are interspersed by abnormally long periods of latency (cold streaks) where authors are less likely than random to work in a given year if they did not work the year before ($\tau > 1$).

Furthermore, to appropriately compare deviations from the null model for different waiting times, in panel (b) of Figure 3 we plot the relative difference (in percentage) $[P(\tau) - P_{\text{null}}(\tau)] \cdot 100/P_{\text{null}}(\tau)$. We find a substantial difference between actors and actresses: while deviation from the null model decays

for larger waiting times τ in the case of actors, for actresses this relative deviation is maintained, pointing to a longer memory kernel, in turn suggesting that having a period of latency is overall more detrimental for actresses than for actors.

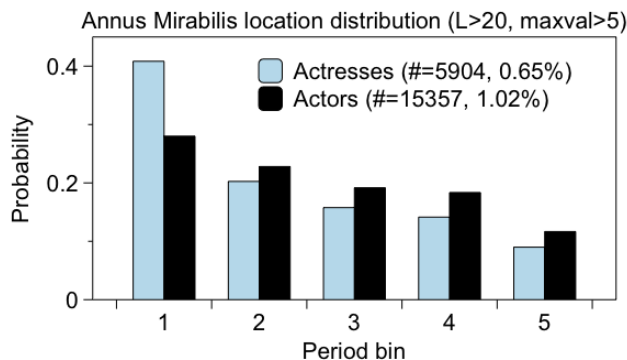


Figure 4: **Annus mirabilis tends to occur sooner rather than later.** Position of AM within an actor or actress’s career, where the career length is binned into 5 bins in every case, to be able to compare profiles of different career lengths. We systematically find that the most probable location of the annus mirabilis is towards the beginning of a career, although this effect is considerably more acute in the case of actresses.

2.5 Predicting the annus mirabilis

It has recently been found that the most impactful publication that a scientist will produce is equally likely to occur at any stage of their career (*I*). Here we explore a related question in the context of actors and actresses. Instead of impact, the indicator of success under study is productivity, as measured by the number of credited works in IMDb. We concentrate on actors and actresses with working lives extending beyond $L = 20$ years. We restrict our reported results to those cases where there were at least 5 credited jobs in the annus mirabilis (AM), although other thresholds do produce qualitatively similar results. The subset of actors with $L > 20$ and more than 5 acting jobs in the AM consists of 15357 actors (1.02%) and 5904 actresses (0.65%). The large gender difference indicates that actors tend to have more acting jobs than actresses.

In Figure 4 we plot the probability with which the AM will occur at each point within an actor or actress’s career. To be able to compare these probabilities over careers of varying lengths, we have broken up each actor’s time series of L years respectively into 5 bins (other segmentations produce qualitatively similar results). The plots consistently indicate that the most probable location of the annus mirabilis is towards the beginning of a career. Although the results are qualitatively similar for male and female actors, this bias is much more pronounced in the case of actresses, further confirming the gender difference previously observed.

To study whether one can detect the imminent appearance of an actor’s annus mirabilis we have analysed, for both actors and actresses, the average number of acting jobs before and after the AM. In order to do this consistently, we initially perform a translation $k \mapsto \kappa$ that aligns all profile sequences,

so that the annus mirabilis $k = y^*$ all occur at $\kappa = 0$. We then define:

$$\xi(\kappa) = \frac{1}{|\mathcal{A}|} \sum_{i=1}^{|\mathcal{A}|} w_{y^*+\kappa}^{(i)},$$

where κ is the offset from the annus mirabilis and $|\mathcal{A}|$ is the size of the set of actors/actresses for which there exists a profile sequence with an input at offset κ . In Figure 5 we plot $\xi(\kappa)$, showing that, on average, there is a clear increase in the number of jobs preceding the AM and a clear decrease immediately afterward. This pattern is absent in the corresponding null models obtained by shuffling the profile sequences (red bars).

It is interesting to note that similar patterns have been observed before in the context of scientific productivity, although recent research challenges this paradigm (24). As a matter of fact, in (24) the authors leveraged the observed shapes of scientific productivity profiles and followed an unsupervised learning approach to cluster different types of careers. Here, instead, we shall follow a supervised learning approach and will now show how the observed patterns can indeed be exploited to build a method for the early prediction of the annus mirabilis.

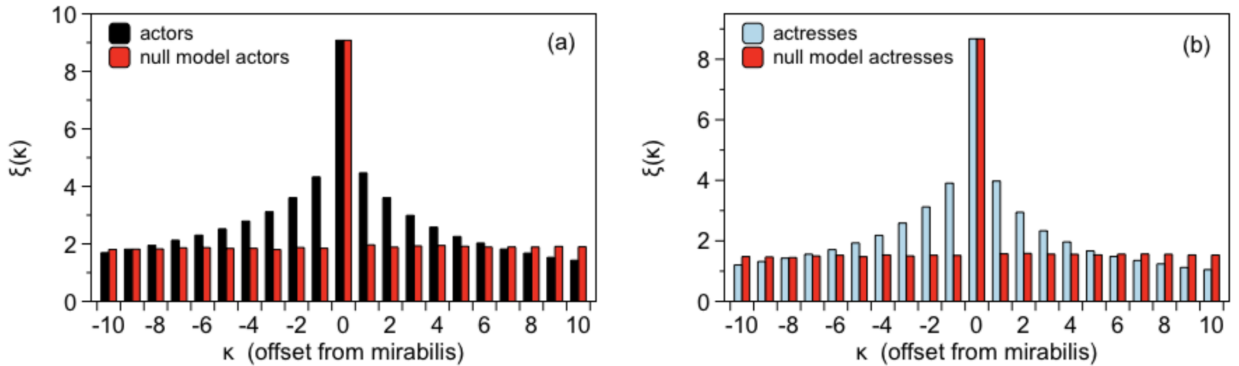


Figure 5: **The annus mirabilis is predictable.** The total number of acting jobs, $\xi(\kappa)$, averaged over all (a) actors and (b) actresses, is reported as a function of the number of years κ after or before the annus mirabilis. Only actors and actresses with a career lasting more than $L = 20$ years and annus mirabilis with $w > 5$ acting jobs have been selected. In both cases, we observe a clear non-monotonic pattern, indicating that the annus mirabilis is either approaching or has just passed. For comparison, we report in red the results obtained for a null model where the profile sequences of all actors and actresses have been shuffled. No pattern emerges in that case.

Based on our observed distribution of jobs surrounding the annus mirabilis we initially propose a naive early-warning criterion: if the career sequence is non-monotonic around a value of k , i.e. if $w_k > w_{k-1}$ and $w_{k+1} < w_k$, then the year k is a good candidate for the annus mirabilis. With this criterion in mind, one could ask the following question: given a sample of an actor or actress's profile sequence, can we tell whether the annus mirabilis has already passed or not? Mathematically, the question above can be formalised as follows: given a career sequence $(w_k)_{k=1}^L$ such that the maximal total productivity occurs at time $k = y^*$, consider a truncated sequence $\bar{w}_k = (w_k)_{k=1}^{T \leq L}$. We

now wish to know if we can accurately assess whether $y^* \in \{1, \dots, T\}$ using only \bar{w}_k . This forms a binary classification problem, in which $\bar{w}_k \in \mathcal{C}_1$ if $y^* \notin \{1, \dots, T\}$ and $\bar{w}_k \in \mathcal{C}_2$ otherwise. Our naive criterion, as illustrated above, readily provides the heuristic: $\bar{w}_k \in \mathcal{C}_1$ if \bar{w}_k is monotonic, and $\bar{w}_k \in \mathcal{C}_2$ if not. When this method is tested on an appropriately generated set \mathcal{W} of truncated sequences (see Supplementary Note 3 for details) we find that it is correct $\sim 69.2\%$ of the times for actors, and $\sim 75.0\%$ of the times for actresses. This model now forms a benchmark against which we will test a more refined approach. The idea is to relax our classification method by introducing some parameters which allow for deviation from the rigid heuristic, then train those parameters on some subset $\mathcal{T} \subsetneq \mathcal{W}$, and subsequently test the trained model on the test set $\mathcal{W} \setminus \mathcal{T}$. To do this let us first define the function

$$D(\bar{w}_k) = - \sum_{y=1}^{T-1} \min(0, \bar{w}_{y+1} - \bar{w}_y). \quad (1)$$

At each year k the contribution to D from that year is zero if the total productivity in the subsequent year is larger. This means that for a monotonically increasing sequence \bar{w}_k , $D(\bar{w}_k) = 0$. If productivity decreases from year k to $k + 1$, then D will increase by a corresponding amount.

$D(\bar{w}_k)$ effectively measures how far the sequence \bar{w}_k is from being monotonically increasing, thus we can use it to relax our naive heuristic by defining some threshold d such that the decision rule $C(\bar{w}_k, d)$ becomes

$$C(\bar{w}_k, d) = \begin{cases} \mathcal{C}_1 & \text{if } D(\bar{w}_k) < d \\ \mathcal{C}_2 & \text{if } D(\bar{w}_k) \geq d. \end{cases}$$

This new classifier is more flexible than the naive heuristic as we have introduced a parameter d which can now be optimised (trained) as follows: if we denote $C^*(\bar{w}_k)$ as the true class of the sequence \bar{w}_k , then the optimal value of the parameter d^* is the value of d that minimises the following loss function

$$L(\mathcal{T}, d) = - \sum_{\mathcal{T}} \delta(C(\bar{w}_k, d), C^*(\bar{w}_k)). \quad (2)$$

Where $\delta(X, Y)$ yields one if $X = Y$ and 0 otherwise. This value for d^* is then used to classify the remaining sequences in $\mathcal{W} \setminus \mathcal{T}$. The results of this testing on both actors and actresses can be partially summarised by the two confusion matrices CO_m (for actors) and CO_f (for actresses):

$$\text{CO}_m = \begin{bmatrix} 33775 & 5659 \\ 10771 & 52000 \end{bmatrix}, \quad \text{CO}_f = \begin{bmatrix} 12549 & 2593 \\ 3596 & 26682 \end{bmatrix}$$

The classical metrics used to assess the performance of the classifier, namely accuracy, precision, recall and the F1 score, are summarised in Table 1. We find that the accuracies of the prediction are 84% and 86% respectively, i.e. $\sim 10\%$ higher than those obtained using a naive heuristic.

To round off, we have further explored the nature of the $\approx 15\%$ of samples which are misclassified (see Supplementary Note 3 for details). We found that false negatives (samples for which the annus mirabilis is wrongly predicted to be still yet to come) arise due to the conservative nature of the prediction model, hence more refined versions of the prediction model might yield even better prediction results (Supplementary Figure 3). Conversely, we find that false positives –where the annus mirabilis

Quantity	Actors	Actresses
Total C_1	44652	16145
Total C_2	57553	29275
Accuracy	0.8405	0.8637
Precision	0.8608	0.8287
Recall	0.7575	0.7773
F1 score	0.8058	0.8021

Table 1: Performance metrics (accuracy, precision, recall and F1 score) of the proposed classification method for the prediction of the annus mirabilis.

is wrongly predicted to have passed—are usually related to actors and actresses experiencing a comeback at a later stage of their careers (see the left panel of Fig 6 for an example). Interestingly, the positions of these late bursts of activity seem to be fundamentally difficult to predict (right panel of the same figure).

3 Discussion

In this work we have made use of the vast quantity of data presented by IMDb to explore, analyse and predict success on the silver screen. By studying the careers of 1,512,472 actors and 896,029 actresses from 1888 up to 2016, we have uncovered a number of distinctive patterns which include an endemic scarcity of resources, a rich-get-richer mechanism of job assignment, the onset of hot and cold streaks of productivity (11) and an annus mirabilis which can indeed be predicted. Such patterns—which we show to systematically differ for actors and actresses, suggesting strong evidence of gender bias (37)—not only allow us to identify qualities of individual actors or actresses working lives, but also to gain a deeper insight into the mechanisms by which jobs are themselves assigned, where high productivity is not necessarily based on merit and is likely to be a network effect (34–36). Based on our findings, we have then constructed a statistical learning model that predicts with up to 85% accuracy whether an actor or actress is likely to have a brighter future, or if the best days are, unfortunately, behind them. While we expect refined versions of the prediction model to give even higher accuracy, it is worth noting that actors with long latency periods who then experience late comebacks are rare but intrinsically difficult to predict.

We hope that the methods presented and the results obtained will contribute to the new science of success (35). Given the scope of our findings across the industry, we also wish that our article will be of interest to those working in show business.

Data Availability

Data and codes are available upon request, or can be accessed at DOI 10.17605/OSF.IO/NDTA3

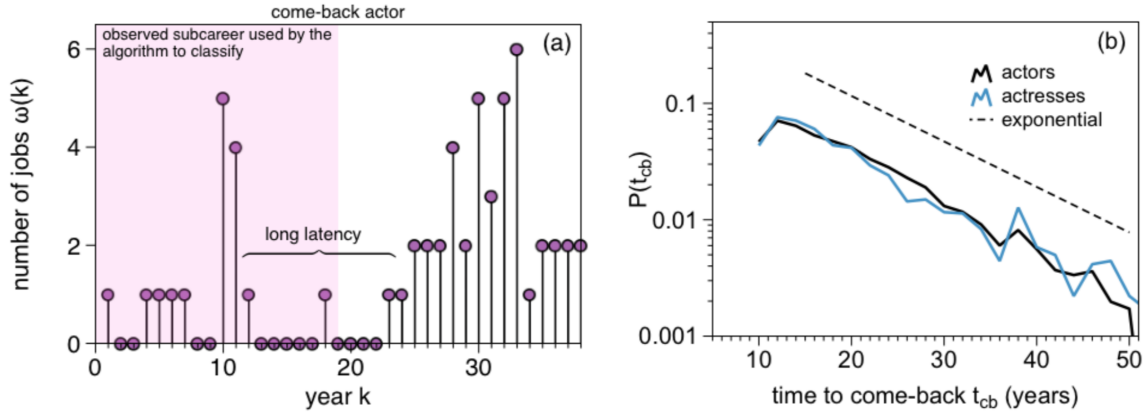


Figure 6: **Comebacks of actors are unpredictable.** (a) Typical profile sequence of an actor exhibiting a comeback after a long period of latency. Such cases might lead to misclassification when the subcareer fed to the prediction algorithm (highlighted in pink) captures a long latency period: the prediction algorithm wrongly classifies the pink sequence as one where the annus mirabilis has passed. (b) Semi-log probability distribution of the estimated time lapse from the (wrongly estimated) annus mirabilis to the true one (i.e. the time t_{cb} to come-back for actors with profile sequences such as the one in the left panel), for those misclassified samples where the algorithm wrongly predicts that the annus mirabilis had already passed (a linear binning has been applied to the data). Modelling the position of the secondary peak (comeback burst) as a random variable, the fact that $P(t_{cb})$ decays exponentially, suggests that this random variable is memoryless (Poisson process), i.e. the comeback burst is intrinsically unpredictable.

Acknowledgments

LL acknowledges support from EPSRC grant EP/P01660X/1. VL acknowledges support from EPSRC Grant EP/N013492/1.

Author contributions

LL, VL and OW designed the study. OW and LL performed the data analysis. All authors interpreted results and wrote the paper.

Competing interests

The authors declare no competing interests.

References

1. R. Sinatra, D. Wang, P. Deville, C. Song, and A-L Barabasi, Quantifying the evolution of individual scientific impact, *Science* 354, 3612 (2016).

2. David W. Galenson. Quantifying artistic success: Ranking french painters and paintings from impressionism to cubism. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 35(1):5–19, 2002.
3. Simone Brands, Stephen J. Brown, and David R. Gallagher. Portfolio concentration and investment manager performance. *International Review of Finance*, 5(3-4):149–174, 2006.
4. Thomas Gilovich, Robert Vallone, and Amos Tversky. The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17(3):295 – 314, 1985.
5. Matthew Rabin and Dimitri Vayanos. The gambler’s and hot-hand fallacies: Theory and applications. *The Review of Economic Studies*, 77(2):730–778, 2010.
6. S. Fortunato Growing time lag threatens Nobels. *Nature* 508, 186 (2014).
7. Athen Ma, Raúl J. Mondragón, and Vito Latora. Anatomy of funded research in science. *Proceedings of the National Academy of Sciences*, 112(48):14760–14765, 2015.
8. James H Love and Stephen Roper. Location and network effects on innovation success: evidence for uk, german and irish manufacturing plants. *Research policy*, 30(4):643–661, 2001.
9. Robert K. Merton. The Matthew effect in science. *Science*, 159(3810):56–63 (1968).
10. Alexander M. Petersen, Woo-Sung Jung, Jae-Suk Yang, and H. Eugene Stanley. Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proceedings of the National Academy of Sciences*, 108(1):18–23, 2011.
11. L. Liu, Y. Wang, R. Sinatra, C. Lee Giles, C. Song and D. Wang, Hot streaks in artistic, cultural, and scientific careers, *Nature* 559, 396-399 (2018).
12. J. E. Hirsch. An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005.
13. Aaron Kozbelt. One-hit wonders in classical music: Evidence and (partial) explanations for an early career peak. *Creativity Research Journal*, 20(2):179–195, 2008.
14. Dean Keith Simonton. Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological review*, 104(1):66, 1997.
15. Harvey Christian Lehman. *Age and achievement*, volume 4970. Princeton University Press, 1953.
16. Dean K Simonton. Age and outstanding achievement: What do we know after a century of research? *Psychological bulletin*, 104(2):251, 1988.
17. Andreas Spitz and Emőke-Ágnes Horvát. Measuring long-term impact based on network centrality: Unraveling cinematic citations. *PloS one*, 9(10):e108857, 2014.
18. Daniel E Acuna, Stefano Allesina, and Konrad P Kording. Future impact: Predicting scientific success. *Nature*, 489(7415):201, 2012.

19. Orion Penner, Raj K Pan, Alexander M Petersen, Kimmo Kaski, and Santo Fortunato. On the predictability of future impact in science. *Scientific reports*, 3:3052, 2013.
20. Gerda Gemser, Martine Van Oostrum, and Mark AAM Leenders. The impact of film reviews on the box office performance of art house versus mainstream motion pictures. *Journal of Cultural Economics*, 31(1):43–63, 2007.
21. Márton Mestyán, Taha Yasseri, and János Kertész. Early prediction of movie box office success based on wikipedia activity big data. *PloS one*, 8(8):e71226, 2013.
22. Iain Pardoe and Dean K Simonton. Applying discrete choice models to predict academy award winners. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2):375–394, 2008.
23. www.imdb.com
24. S.F. Way, A.C. Morgan, A. Clauset, and D.B. Larremore, The misleading narrative of the canonical faculty productivity trajectory, *Proc. Natl. Acad. Sci.* 114, 44 (2017).
25. Consistent percentages obtained via large surveys have been reported in several countries including US, UK or Spain see: N.Clark, *Just one actor in 50 makes more than 20,000 per year; survey shows* (The Independent, 28th May 2014, online version available at <https://www.independent.co.uk>); *Estudio y Diagnóstico sobre la situación sociolaboral de actores y bailarines en España*, Fundación AISGE, 2016, available at <https://www.aisge.es/media/multimedia/ficheros/618.pdf>; B. McMahon, *Unemployment is a lifestyle for actors, and not too many others*, Huffpost, 5 March 2012., available at <https://www.huffingtonpost.com/brendan-mcmahon>
26. B. Yucesoy and A-L Barabasi, Untangling performance from success, *EPJ Data Science* (2016) 5:17
27. Dale T. Mortensen. Chapter 15 job search and labor market analysis. volume 2 of *Handbook of Labor Economics*, pages 849 – 919. Elsevier, 1986.
28. M. Latapy, C. Magnien, N. Del Vecchio, Basic notions for the analysis of large two-mode networks, *Social Networks* 30 (2008) 31–48.
29. L.A. Adamic and B.A. Huberman, Zipf’s law and the Internet’, *Glottometrics* 3 (2002) pp.143–150
30. A-L Barabasi and R. Albert. Emergence of scaling in random networks, *Science* 286 (1999), pp.509–512.
31. Bela Bollobas and Oliver Riordan. Mathematical results on scale-free random graphs. In Stefan Bornholdt and Hans Georg Schuster, editors, *Handbook of Graphs and Networks*, pages 1–34. John Wiley & Sons, 2005.
32. V. Latora, V. Nicosia, G. Russo, *Complex networks: principles, methods and applications* (Cambridge University Press 2017)

33. D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World* Chapter 18 (Cambridge University Press, 2010)
34. M. Salganik, P. Dodds, and D. Watts. Experimental study of inequality and unpredictability in an artificial cultural market, *Science* 311 (2006), pp.854–856.
35. A.L. Barabasi, *The Formula: The Universal Laws of Success* (Little Brown and Company, New York, 2018)
36. Fraiberger, S. P., Sinatra, R., Resch, M., Riedl, C., and Barabási, A. L. (2018). Quantifying reputation and success in art. *Science*, 362, 6416 (2018), 825-829.
37. M. Lutter, Do women suffer from network closure? The moderating effect of social capital on gender inequality in a project-based labor market, 1929 to 2010, *American Sociological Review* 80, 2 (2015): 329-358.
38. P. Bak, K. Christensen, L. Danon, T. Scanlon, Unified scaling law for earthquakes, *Phys. Rev. Lett.* 88, 178501 (2002).
39. A. Corral, Long-Term Clustering, Scaling, and Universality in the Temporal Occurrence of Earthquakes, *Phys. Rev. Lett.* 92, 108501 (2004).
40. K.P. Burnham, D.R. Anderson, *Model Selection and Multimodel Inference: A practical information-theoretic approach* (2nd ed.) (Springer-Verlag, 2002).