Supporting Information

Way et al. 10.1073/pnas.1702121114

Collection of CV Data

Performing the DBLP coverage analysis and adjustment discussed in General Trends in Productivity Data required a benchmark dataset with complete coverage of the publication histories of a representative subset of researchers. This section describes the relevant details of the collection of that benchmark dataset. We manually extracted lists of publication dates from publicly available CVs belonging to a random 10% of the N = 1091 researchers with career lengths between 10 years and 25 years and having publications in at least 3 distinct years. This last condition ensures that the piecewise linear model can be fitted to the individual’s trajectory and excludes just 32 researchers from our analysis. Because of the high diversity of productivity trajectories, we chose 10% of individuals, uniformly at random, from each of the quadrants designated by the signs of the two slope parameters, \( m_1 \) and \( m_2 \), as shown in Fig. 4. Specifically, names of researchers from each quadrant were randomly shuffled and then collected, in order, until reaching 10% of the total. However, individuals for whom a CV could not be found or whose publicly available CV was last updated before 2011 were skipped, and other faculty were randomly selected in their place. Success rates for this exercise ranged between 66.3% and 86.6%, measured as the number of successfully extracted publication lists vs. the total number of attempts. The majority of these failures were due to researchers having out-of-date CVs. Future studies should consider whether such partial records of researcher productivity are sufficient for analysis, as their inclusion would greatly improve success rates during collection.

General Trends in Productivity Data

Meaningful trends in publication rates over a career can be confidently identified from raw publication counts only if two conditions are met. First, raw publication counts must be exhaustive, containing all peer-reviewed publications. Second, field-wide publication rates must be stationary over time. Due to the facts that DBLP data do not satisfy the former, and that computer science as a field does not satisfy the latter, raw publication counts recorded in the DBLP dataset must be adjusted to compensate before they can be analyzed. This section explains the details of two compensatory adjustments to DBLP data. We justify the first adjustment by providing a detailed analysis of the time-varying fraction of publications covered by the DBLP dataset, anchored by a hand-collected benchmark CV dataset (Collection of CV Data). We then justify the second adjustment by identifying a clear and significant overall growth in publication rates over 40 years of computer science publication data. We conclude by discussing several possible explanations for why researcher productivity increases over time.

DBLP has indexed the overwhelming majority of current computer science publications, including peer-reviewed conferences and journals. However, while excellent today, this coverage has increased systematically over time, meaning that DBLP coverage is less complete for older publications and faculty. To quantify trends in the time-varying coverage of our DBLP data, we hand collected the CVs of 109 faculty, representing 10% of individuals whose trajectories are shown in Fig. 4, providing a set of benchmark publication lists (Collection of CV Data).

For each year of our dataset, we compared the number of publications listed on individuals’ CVs to the number of publications listed on their corresponding DBLP profiles, selecting only peer-reviewed conference and journal publications from CVs. Comparing DBLP data with CV benchmarks, two sets of counts reveal that DBLP coverage has increased linearly from around 55% in the 1980s to over 85% in 2011 (Fig. S1). DBLP coverage has grown at a rate of \(~1.06\%\) of additional coverage per year, with a 95% confidence interval indicating that this rate falls between 0.8% and 1.4%. Because the ratio of DBLP publications \( y_{DBLP} \) to CV publications \( y_{CV} \) is well described by the line

\[
\frac{y_{DBLP}(t)}{y_{CV}(t)} = m_0 t + b_0, \quad \text{[S1]}
\]

we use Eq. S1 to convert all nonbenchmarked DBLP publication counts to CV-equivalent publication counts, with estimated parameters of \( m_0 = 0.010588 \) and \( b_0 = -20.434804 \).

After linearly adjusting all raw publication counts to correct for the expected DBLP coverage in a given year (Eq. S1), we analyzed how individual researcher productivity has changed over the years spanned by our dataset. Due to the fact that we extracted and adjusted DBLP records for only authors in the faculty hiring dataset (1, 26), a straightforward analysis of the number of per-person publications for each calendar year would feature a different mixture of career ages in each year. For example, adjusted publication counts from the 1970s would include only early-career researchers; late-career researchers in the 1970s retired long before our dataset was collected. Because the main text of this paper reveals systematic trends in productivity by career age, this straightforward counting technique would introduce bias.

To quantify the expansion of publication rates over time, without introducing career-age bias, we selected “indicator” career ages at which to measure productivity and compared how productivity at specific points in the academic career has changed over time. The trend in publication growth is consistently positive for all indicator career ages, and, as in the main text, publication rates after year 5 tend to be higher than publication rates in the first 4 years. When these publication rates were normalized by their 2011 values, all indicator sets collapsed onto a common growth line (Fig. S2). Thus, while the indicator sets of \{0, 1, 2, 3, 4\}, \{5\}, and \{5, 10, 15\} revealed that productivity grows at rate between 0.84 and 1.48 additional papers per person per decade, their rates of growth are directly proportional to 2011 productivity, meaning that the shape of the canonical trajectory has not changed over time, but has simply expanded proportionally.

The relative slopes of publication rate expansion are consistent with the relative publication rates in the canonical “average” productivity trajectory (Fig. 2). Indeed, growth in researchers’ first 5 years of productivity is relatively modest, which is expected since researchers, both historically and more recently, spend these years building their research programs by applying for funding and recruiting graduate students and postdoctoral researchers. On the other hand, productivity in year 5—the year of or immediately preceding tenure evaluations at most institutions and, perhaps not coincidentally, the modal year of peak researcher productivity—grows at a faster rate of 1.48 additional papers per person per decade. These rates are both similar comparing years 5, 10, and 15, which describes changes across a larger window of career productivity.

The relationship between 2011-equivalent publications and past publications is consistently linear over the time spanned by our dataset (Fig. S2) and is modeled well by...
We therefore used Eq. S2 to convert all CV-equivalent publication counts to 2011-equivalent publication counts, with estimated parameters \( \hat{m}_0 = 0.131873 \) and \( \hat{b}_0 = -258.286620 \). Thus, we applied the two transformations of Eq. S1 (Fig. S1) and Eq. S2 (Fig. S2) in series to publication data to produce the adjusted publication counts used in the main text, unless otherwise specified; benchmark CV data were used for the 109 individuals for whom we collected them, and for those individuals, only Eq. S2 was applied. Note that while model-based correction can adjust for trends in publication rates, alternative model-free approaches have been used by other researchers, which convert publication rates to annual publication rate \( z \) scores, e.g., ref. 36.

In the main text, we noted that, after applying these two linear adjustments, the median number of early-career publications per person per institution increases over time and correlates with prestige. Fig. S3 illustrates this trend, stratifying individuals into five levels of prestige and revealing that production growth rates between higher- and lower-prestige departments have widened slightly but significantly over time (\( P < 0.05 \), two-tailed \( t \) test). This trend, observed for early-career publications (i.e., publications within the first 10 years of a career, for individuals with careers of 10 years or longer), is no different from the trend for all posthire publications and all researchers (Fig. S4). Median lifetime career productivity correlates significantly with prestige, and, as in early-career productivity, public and private institutions are similarly affected by this relationship. Using an ordinary least-squares regression of productivity vs. prestige including dummy and interaction terms for public/private status, we found that the relationship between productivity and prestige is not significantly affected by public/private status (\( P > 0.05 \), \( t \) test, for both public/private dummy and public/private-prestige interaction).

While production rates have increased steadily over time and for all levels of prestige, we find that the imbalance in research production has decreased in recent years. Fig. S5 illustrates this inequality across researchers and time by showing the fraction \( Y \) of all publications in our sample that were produced by the most productive fraction \( X \) of all faculty (a Lorenz curve), for faculty first hired in each of the four decades that our data span and restricting analysis to only publications produced in the first 5 years of an individual’s career. As referenced in the main text, the Gini coefficients for productivity imbalance have declined, from 0.62 in the 1970s to 0.40 in the 2000s. Many factors could potentially drive such a shift toward more balanced research production in science (including, for example, technological advancements and corresponding declines in research costs that may have leveled the playing field for researchers at less-prestigious universities), and we welcome future studies that explore this shift in more detail.

Finally, several possibilities exist that might explain why researchers are becoming more productive. First, the average number of coauthors per publication has steadily increased over time, allowing researchers to work on a larger number of projects. Second, the number of publication venues has also grown, providing more outlets for researchers’ work and potentially facilitating more specialized communities with faster peer review. Third, technological advances including improvements in computer architecture have benefitted researchers universally, increasing the speed at which results are both generated and published. Finally, perhaps the perception of what constitutes the minimum publishable unit of research has changed over time, resulting in a larger number of shorter, more narrowly focused publications in recent years.

**Modeling Framework**

Eq. 1 is the simple model used in the main text to parameterize the adjusted publication counts over the course of a career. Reproduced below, it consists of two lines with slopes \( m_1 \) and \( m_2 \) that intersect at time \( t^* \).

\[
f(t) = \begin{cases} 
  b + m_1 t & 0 \leq t \leq t^* \\
  b + m_1 t^* + m_2 (t - t^*) & t > t^*
\end{cases}
\]  

S3

In this article, we fit Eq. S3 to adjusted count data, using least squares. However, there exist other regression frameworks that correspond to generative models for time series data. In this section, we discuss some of these alternatives.

**Adjusting Models Instead of Counts.** Although publication time series are naturally count data, a regression framework that is naturally suited for counts, such as Poisson or negative-binomial regression, is not advised. Directly fitting raw counts using a Poisson or negative-binomial model would neglect the adjustments for both the coverage of DBLP and the time-varying changes in publication rates (General Trends in Productivity Data). On the other hand, adjusted publications are nonintegers, rendering them inappropriate for count regressions. However, there are alternatives that would allow for both count regressions and the adjustments in General Trends in Productivity Data. These adjustments come with a price, however, due to the assumptions and free parameters that they introduce.

One alternative solution to fitting a model to adjusted publication data is to fit an adjusted linear model to raw publication data. In other words, adjust the model instead of the data. Due to the fact that this approach would preserve the data as counts, this model would be amenable to Poisson and negative-binomial regression frameworks. Adjusted publications \( y_{2011} \) are related to raw publications by the adjustment

\[
y_{2011}(t) = y_{DBLP}(t) \frac{1}{\hat{m}_{a} t + \hat{b}_{a}} \frac{1}{\hat{m}_{b} t + \hat{b}_{b}},
\]  

S4

where \( m_a \) and \( m_b \) are slopes and \( b_a \) and \( b_b \) are intercepts of the linear adjustments for DBLP coverage and publication expansion, respectively, and hats indicate that variables have been estimated from data (General Trends in Productivity Data). Applying this adjustment to the model \( f \), which is to hold for adjusted publications, we get

\[
f_{DBLP}(t) = (\hat{m}_{a} t + \hat{b}_{a}) (\hat{m}_{b} t + \hat{b}_{b}) f(t - t_0).
\]  

S5

where \( t_0 \) is the initial year of a particular faculty member’s career, and \( t \) is the calendar year.

We now turn to details of Poisson and negative-binomial frameworks for fitting the adjusted model Eq. S5 and discuss the assumptions and parameters that they introduce.

**Poisson Model.** Consider a Poisson fit of Eq. S5 to a set of data given by \( \{t_i, y_i\} \). To simplify, let us be explicit about the dependence of \( f_{DBLP} \) on the four parameters, \( m_1, m_2, \hat{b}_{a}, \) and \( \hat{b}_{b} \), which we collectively refer to as \( \theta \).

\[
f_{DBLP}(t; \theta) = q(t) f(t - t_0; \theta),
\]  

where we have made clear that \( q(t) = (\hat{m}_{a} t + \hat{b}_{a}) (\hat{m}_{b} t + \hat{b}_{b}) \) does not depend on the model parameters \( \theta \). The likelihood is then

\[
P(\{t_i, y_i\}; \theta) = \prod_i e^{-q(t_i) f(t_i - t_0; \theta)} \left[ q(t_i) f(t_i - t_0; \theta) \right]^{y_i} \frac{1}{y_i!}.
\]  

S6
Rather than maximizing $P$, we maximize $\log P$. Taking the natural log of both sides, we get

$$\log P(t_i, y_i | \theta) = \sum_i \left\{ -q(t_i)f(t_i - t_0; \theta) + y_i \log q(t_i) + \log f(t_i - t_0; \theta) \right\} - y_i \log y_i$$

Note that the terms $-\log y_i!$ and $y_i \log q(t_i)$ do not depend on the parameters $\theta$, so they affect the value of the maximum but not its location in parameter space. Dropping them yields a log-likelihood score $L$ of

$$L(t_i, y_i | \theta) = \sum_i -q(t_i)f(t_i - t_0; \theta) + y_i \log f(t_i - t_0; \theta).$$

Note that for any trajectory, $q(t_i)$ can be precomputed and does not depend on the parameters $\theta$. Thus, fitting the 2011-equivalent Poisson model requires that we maximize Eq. S8 with respect to $m_1, m_2, b$, and $\tau$. This equation must be maximized numerically.

While this adjusted Poisson model is attractive because it naturally fits count data, it imposes assumptions on the data-generating process that are not justified empirically. Namely, the variance and mean of a Poisson distribution are equal, meaning that the Poisson regression expects the same of the data it explains.

**Negative-Binomial Model.** The Poisson model above enforces the constraint that the mean is equal to the variance. However, there is no indication that the data support this assumption so we introduce the standard alternative, the negative-binomial model. This model requires both a mean $\mu$ and a heterogeneity parameter $\zeta$, such that the probability of a single observation $y$ is

$$P(y) = \frac{\Gamma(y + \frac{1}{\zeta})}{\Gamma(y + 1)\Gamma(\frac{1}{\zeta})} \left( \frac{1}{1 + \zeta \mu} \right)^{\frac{1}{\zeta}} \left( \frac{\zeta \mu}{1 + \zeta \mu} \right)^y.$$  \[[S9]\]

As in the Poisson regression, we once more parameterize the mean, using the piecewise linear model as $\mu = \mu(t_i) = q(t_i)f(t_i - t_0)$. However, we must also introduce a model for $\zeta(t_i)$.

The easiest way forward, mathematically, is to set $\zeta(t) = \zeta$. This assumes equal heterogeneity around the expected value $\mu(t_i)$ for all time points in a career $t_i$. Note that this assumption decouples the heterogeneity from the mean, while under the Poisson model they are directly coupled. One might think of the Poisson model therefore as fitting the parameters $\theta$ to both the trend and the fluctuations together. The fixed-$\zeta$ negative-binomial model, on the other hand, fits the parameters $\theta$ to the trend and uses a fixed $\zeta$ to accommodate all fluctuations. In this sense this negative-binomial approach is more flexible and uses an additional parameter to gain that flexibility.

The $\zeta(t_i) = \zeta$ assumption results in a log probability of

$$\log P(t_i, y_i | \theta, \zeta) = \sum_{i=1}^T \left\{ \log \Gamma \left( y_i + \frac{1}{\zeta} \right) - \log \Gamma(y + 1) - \log \Gamma \left( \frac{1}{\zeta} \right) - \frac{1}{\zeta} \log [1 + \zeta q(t_i)f(t_i - t_0)] ight\}$$

and we note that $\sum_{i=1}^T \log \Gamma \left( \frac{1}{\zeta} \right) = T \log \Gamma \left( \frac{1}{\zeta} \right)$ and that both $\sum_{i=1}^T \log \Gamma(y_i + 1)$ and $\sum_{i=1}^T y_i \log q(t_i)$ are constants that do not depend on either $\theta$ or $\zeta$, allowing us to write a log-likelihood score of

$$L(t_i, y_i | \theta, \zeta) = -T \log \Gamma \left( \frac{1}{\zeta} \right) + \sum_{i=1}^T \left\{ \log \Gamma \left( y_i + \frac{1}{\zeta} \right) - \frac{1}{\zeta} \log [1 + \zeta q(t_i)f(t_i - t_0)] + y_i \log [\zeta q(t_i)f(t_i - t_0; \theta)] - \log [1 + \zeta q(t_i)f(t_i - t_0; \theta)] \right\}.$$  \[[S10]\]

Progress here, however, is obstructed by the difficulties of taking derivatives of Gamma functions. Thus, Eq. S11 must be optimized numerically over the parameters $\theta$ and $\zeta$.

While these calculations may be helpful in seeding a path forward in future work, it is important to note that the fixed-$\zeta$ negative-binomial model also makes strong assumptions about the generative process that created the data. Indeed, the assumption that fluctuations are uniform over an entire career is strong and is not justified by data. One could also avoid this assumption, but this introduces additional problems, which we now discuss.

The temptation to let each point $t_i$ have a parameterized value of $\mu(t_i)$ and a free parameter of $\zeta(t_i)$ results in overfitting. Note that this would allow each point in the time series $(t_i, y_i)$ to be fitted by a negative-binomial distribution with a mean given by Eq. S5 and an arbitrarily large or small $\zeta(t_i)$, resulting in dramatic overfitting. This approach therefore makes few assumptions, but provides little value to the modeler.

A middle ground between fixed $\zeta$ and unrestricted $\zeta(t_i)$ would be to parameterize $\zeta(t_i)$, using a lower-dimensional model. Using the same model for $\zeta(t_i)$ as we used for $\mu(t_i)$ would be similar, in principle, to the Poisson regression. Using a different model for $\zeta(t_i)$ is an option, but would require, again, a deep focus on the underlying mechanisms hypothesized to explain fluctuations in productivity.

**Modeling Outlook.** Generative models for productivity trajectories would be enormously valuable. In this section, we emphasize the assumptions made by the generative models underlying various regression frameworks. In particular, we derive models that are able to be fitted directly to raw count data by including the inverse of the adjustment derived in General Trends in Productivity Data, a quadratic term referred to as $q(t)$.

In terms of impacts, fitting the Poisson model and the fixed-$\zeta$ negative-binomial model to the trajectories investigated in this paper does affect the parameters of individuals’ trajectories. However, it does not diminish the diversity of trajectories that we observe. Indeed, the example trajectories shown in Fig. 4 are only subtly affected by the use of one type of generative model or another.

**Sensitivity to Timing of Publications.** Publication generally signifies the conclusion of a research project but the exact date when an article is published can depend on many factors, including the availability of reviewers, graduation deadlines for graduate students, delays between acceptance and publishing, and synchronization with conference submission deadlines, as well as nonacademic constraints, such as the impending birth of a child. Each of these factors might advance or delay a publication’s appearance in the literature, and furthermore, the effort associated with each publication may depend on many factors, including the availability of reviewers, graduation deadlines for graduate students, delays between acceptance and publishing, and synchronization with conference submission deadlines, as well as nonacademic constraints, such as the impending birth of a child. Each of these factors might advance or delay a publication’s appearance in the literature, and furthermore, the effort associated with each publication may span weeks, months, or years. As a result, publication years serve as a noisy indicator of when productivity occurs.
To ensure that our findings are not due to coincidence in the timings of researchers’ publications, we examined the sensitivity of our results to the addition of small amounts of noise. For each researcher and for each of their publications, we added noise drawn from a normal distribution ($\mu = 0, \sigma = 0.7413011$) to the publication year and then rounded to the nearest whole year. In expectation, this process leaves one-half of publication years unaffected and shifts 22.8% by 1 year in either direction. 2.1% by 2 years, and 0.04% by 3 years. We repeated this process 200 times for each researcher (examples shown in Fig. S6) and found that the median number of trials in which the trajectory changed shape compared with the noise-free fit (i.e., $m_1$ or $m_2$ changed sign) was 9 (4.5%; Fig. S7).

While the typical individual’s parameters are robust to noise, those individuals whose trajectories featured few publications or whose noise-free model parameters were near zero were far more likely to change shape. In fact, for 10.5% of individuals, noise led to shape change more often than not, i.e., in greater than 50% of noisy repetitions. Therefore, as an additional check, we asked whether the model parameters inferred for researchers’ noise-free trajectories differ significantly from those inferred for their 200 noise-added trajectories. We used Fisher’s method to combine $P$ values for each researcher and found no significant differences in any of the four model parameters ($m_1$, $m_2$, $b$, and $t^*$). Evaluated separately, fewer than 1% of researchers’ inferred model parameters differ significantly under the noise-free and noise-added fits. We conclude that the general shape of productivity trajectories is robust to small differences in publication year.

Having specified a model for noise around each publication date, an individual’s trajectory naturally becomes a distribution of trajectories, which in turn maps into a corresponding distribution in the $m_1 \times m_2$ parameter space. To analyze this distribution, we inferred model parameters for each of a researcher’s 200 noise-added trajectories, and for those trajectories that were more confidently modeled by Eq. 1 (using the AIC with finite-size correction; Model Selection) vs. a straight line, we compiled the distribution depicted in Fig. S8.

This distribution suggests a complementary approach to investigating the universality of the conventional narrative wherein individuals are considered as distributions rather than point estimates. The latter of these approaches (presented in the main text) requires specifying a threshold for stability that determines whether or not an individual can be confidently mapped into a particular region in the space. The former, on the other hand, requires no such distinction and instead sums the independent distributions of individual faculty to quantify the total probability mass in each region. The point estimate approach makes a statement about individuals, while the distribution approach makes a statement about the population. Importantly, both statements agree: ~20% of individuals are firmly mapped to the canonical octant, and ~20% of probability mass maps into the canonical octant.

**Model Selection**

When the complexity of a model exceeds the complexity of the underlying data, some parameters of the model may no longer be interpreted as meaningful. Although the piecewise linear model of Eq. 1 has only four parameters—two slopes $m_1$ and $m_2$, an intercept $b$, and a change point $t^*$—it may nevertheless overfit productivity trajectories that are actually linear. This section provides additional details for our model selection procedures that avoid the overinterpretation of the piecewise change-point parameters $t^*$.

If a publication trajectory is generated by a straight line with added noise, then fitting a piecewise model will result in $m_1$ approximately equal to $m_2$, and the location of the change-point $t^*$ will be arbitrary. We apply model selection to identify individuals with career lengths between 10 years and 25 years ($N = 1,091$) whose productivity trajectories are both stable (Sensitivity to Timing of Publications) and consistently better modeled by the piecewise model Eq. 1 than by a straight line [i.e., ordinary least squares (OLS)]. This filter includes only those trajectories whose change-points $t^*$ can be interpreted with confidence.

To perform model selection, we consider three information theoretic model selection techniques: the AIC, the Akaike information criterion with finite-size correction ($AIC_c$), and the Bayesian information criterion (BIC), defined as

$$AIC = n \log(SSE/n) + 2k$$

$$AIC_c = n \log(SSE/n) + 2k + \frac{2k(k + 1)}{n - k - 1}$$

$$BIC = n \log(SSE/n) + k \log(n),$$

where $n$ is the length of the individual’s career, $k$ is the number of model parameters ($k = 2$ for OLS, and $k = 4$ for Eq. 1), and $SSE$ is the sum of squared errors (differences) between actual and modeled publication counts. The AIC is the least conservative of the three methods, finding that 59.1% ($N = 645$) of individuals are better modeled by Eq. 1 than by OLS regression. By contrast, $AIC_c$ and BIC select only 33.2% ($N = 363$) and 44.4% ($N = 485$) of individuals, respectively. In the main text, we adopt the most conservative approach, $AIC_c$, but the other two methods nevertheless produce qualitatively similar distributions for $t^*$, with the modal year for $t^*$ remaining at year 5.

**Detection of Alphabetized Publication Venues**

Conventions of author order vary widely in computer science. In the first/last convention, first authorship is reserved for the lead author or primary contributor to the study, while last authorship indicates the senior author who oversaw or advised the work. In the alphabetical convention, borrowed from mathematics, a paper’s authors are arranged alphabetically by last name. If the trend from first authorship toward last authorship over a career is to be reliably interpreted (Fig. 8), publications with alphabetical author orders must be discarded. This section explains the methods used to statistically identify and remove publication venues that are highly enriched with alphabetical conventions.

Our approach is to count the number of multiauthor papers in each publication venue with alphabetically ordered authors and compare this count to the number expected by chance. (We note that all single-author papers are ignored in this analysis.) A paper with $M$ authors will list its authors alphabetically by chance with probability $1/M!$. Noting the number of authors of each multiauthor paper published by a particular venue and assuming independence of ordering decisions, we derive an empirical distribution representing the number of coincidentally alphabetized author lists and ask whether the venue adopts the convention significantly more often than would be expected by chance. Additionally, we require that the number of observed alphabetized lists be at least twice the expected value. These two conditions ensure that the alphabetical convention is both significant and widespread in its adoption in a particular venue. We find that 630 of the 5,622 (11.2%) distinct venues in our dataset alphabetize their author lists. These venues account for 27,237 of the 177,437 (15.4%) multiauthor conference or journal publications for which the publication venue is known.

Manually inspecting the list of alphabetized venues reveals that popular theoretical venues like the Symposium on Theory of Computing (STOC), Foundations of Computer Science (FOCS), the Symposium on Theoretical Aspects of Computer Science (STACS), and the Symposium on Discrete Algorithms (SODA) adhere to the alphabetical convention, while the
World-Wide Web Conference (WWW), the Conference on Computer-Supported Cooperative Work and Social Computing (CSCW), the Conference on Knowledge Discovery and Data Mining (KDD), the Conference on Human Factors in Computing Systems (CHI), and the AAAI Conference on Artificial Intelligence (AAAI) do not, matching our expectations.

Least-Squares Fit of \( f(t) \)

A least-squares fit of the continuous piecewise-linear equation \( f(t) \) given in Eq. 1 involves minimizing the sum of squared errors over four parameters, \( m_1, m_2, b, \) and \( t^* \). In this section, we provide details that make this fitting process rapid and maximally accurate.

The model fit consists of two steps. First, we assume that \( t^* \) is fixed and find the optimal values of \( m_1, m_2, \) and \( b \). In the second step, we search for the \( t^* \) whose corresponding optimal parameters provide the best fit. The sum of squared error \( \varepsilon \) is given by

\[
\varepsilon = \frac{1}{2} \sum (m_1 t_i + b - y_i)^2 + \frac{1}{2} \sum' (m_1 t^* + m_2 (t_i - t^*) + b - y_i)^2, \tag{S12}
\]

where \( t^* \) is the change point, \( \sum \) denotes the sum for all \( t_i < t^* \), and \( \sum' \) denotes the sum for all \( t_i \geq t^* \).

In the first step, we imagine \( t^* \) to be fixed and simply take partial derivatives with respect to the three parameters, set each equal to zero, and solve. Setting \( \nabla \varepsilon = 0 \) yields three equations,

\[
\begin{align*}
& m_1 \left[ \sum t_i^2 + \sum t_i' \right] + m_2 \sum t_i' (t_i - t^*) \\
& + b \left[ \sum t_i + \sum t_i' \right] = \sum y_i t_i + \sum y_i t^*, \\
& m_1 \sum t_i' (t_i - t^*) + m_2 \sum (t_i - t^*)^2 \\
& + b \sum (t_i - t^*) = \sum y_i (t_i - t^*), \\
& m_1 \left[ \sum t_i + \sum t_i' \right] + m_2 \sum (t_i - t^*) \\
& + b \left[ \sum 1 + \sum t_i' \right] = \sum y_i + \sum y_i. \tag{S13}
\end{align*}
\]

Thus, for a fixed value of \( t^* \), the above equations provide a linear system of three equations for the three unknowns \( m_1, m_2, \) and \( b \). The system can be solved numerically, using any linear solver.

In the second step, we embed the optimization above within a search for the optimal \( t^* \) value. For each proposed value of \( t^* \), we use Eq. S13 to find the optimal \( m_1, m_2, \) and \( b \) and then compute the associated error using Eq. S12, choosing the \( t^* \) that minimizes \( \varepsilon \). Initially, we propose a coarse grid at the level of \( \Delta t^* = 0.1 \) and then refine the grid by an order of magnitude locally around the best result, repeatedly, until the optimal \( t^* \) is known to single precision. This procedure is fast, and due to the result of Eq. S13, limits numerical search to a one-dimensional line.

![DBLP pubs. / CV pubs.](image1)

**Fig. S1.** DBLP coverage improves for more recent publications. Shown is the fraction of all publications found in DBLP data compared with publication lists extracted from CVs of corresponding researchers, separated by year. Regression of these fractions reveals that DBLP coverage improves by approximately 1.06% each year. The shaded region denotes the 95% confidence interval for the regression. pubs, publications.

![Indicator year(s):](image2)

**Fig. S2.** Individual productivity has increased over time. After adjusting for DBLP coverage (General Trends in Productivity Data), evaluation of average individual performances in select indicator years reveals that researchers have become more productive over time, growing at a rate of approximately one additional paper per decade. Shaded regions denote the 95% confidence interval for each regression. pubs, publications.
Fig. S3. Annual publication rates have grown steadily. For faculty in this study, per-person annual publications have increased over time at a rate of approximately one additional paper every 10 years. This rate of growth affects researchers at all levels of prestige rank (26). Slopes represent least-squares linear regressions, with shaded regions denoting corresponding 95% confidence intervals. Pubs, publications.

Fig. S4. Publications correlate with prestige of using institutions. Circles indicate median number of publications per person per institution for all years posthire, adjusted for growth in publication rates over time and ordered by institutional prestige. Effects of prestige are statistically indistinguishable for private (open circles) and public (solid circles) institutions. Shaded region denotes the 95% confidence interval for least-squares regression. Pubs, publications.

Fig. S5. Production imbalance among faculty. Lorenz curves of adjusted production by in-sample faculty, stratified by decade of first hire (key), show that ~20% of faculty account for half of all publications in the dataset. The Gini coefficient $G$ for each curve is noted in the key; the diagonal line indicates equal production by all faculty.
Fig. S6. Example model fits for noise-free and noise-added publication trajectories. Shown are example publication trajectories (gray lines with circles) with piecewise linear fits (black lines) and 200 fits to noise-added trajectories (orange lines). Trajectories are categorized as stable whenever 75% or more of fits fall within a single quadrant, as indicated in Insets.

Fig. S7. Trajectory shapes are robust to perturbations in publication years. Applying the piecewise linear model to 200 noise-added publication trajectories for each researcher, the median fraction of trials resulting in sign changes of model parameters $m_1$ or $m_2$ compared with noise-free fits is 0.045.
Fig. S8. The distribution of noise-added trajectories matches that of individuals. Shown is the distribution of slope parameters $m_1$ and $m_2$ for 200 noise-added trajectories, for each individual and requiring that each instance is not better modeled as a straight line (AIC with finite-size correction). Counts of noise-added trajectories are tabulated as percentages falling into each quadrant and within the octant corresponding to the conventional narrative (19.7%). Num, number.