Supporting Information for "The Chaperone Effect in Scientific Publishing"

Vedran Sekara et al.

1. Alphabetical null model

We begin by analyzing the behavior of the *alphabetical null model* to which we compare the empirical data in the main text Fig. 2.

Consider the following null model to create an authorship list with m names for k papers: Generate an author name A as a random string. Generate k sets of further random author names, where each set contains $m-1$ co-authors of A, as well as A themselves. We have now two choices of author orderings: alphabetical and random. Let us first consider random orderings.

The probability that A is last author on a given publication is simply $1/m$ as there are m authors on a paper. This is therefore also the probability that A is last author on the first of the k papers. Hence:

$$p(A\,\text{last on first paper, with random ordering}) = \frac{1}{m}$$

The probability that A is never last author in any of the k publications is:

$$p(A\,\text{never last, with random ordering}) = \left(1-\frac{1}{m}\right)^k$$

The probability that A is chaperoned, i.e. a last author on at least one of the $k-1$ publications *after* the first is therefore:

$$p(A\,\text{last after first paper, with random ordering}) = 1-\frac{1}{m}-(1-\frac{1}{m})^k$$

Now let us consider alphabetical orderings. Let us map the near-continuum of alphabetically ordered string to the interval $[0,1]$. The probability that A is last author on a given publication then is:

$$p(A\,\text{last on first paper, with alphabetical ordering}) = \int_0^1 x^{m-1}dx = \frac{1}{m}$$

as for the random case. The probability that A is never last author is any of the k publications is:

$$p(A\,\text{never last, with alphabetical ordering}) = \int_0^1 (1-x^{m-1})^k dx$$

This is a more involved integral. Substitute $q = x^{m-1}$ so that $q^{\frac{1}{m-1}} = x$ and

$$\frac{1}{m-1}q^{\frac{1}{m-1}-1}dq = dx$$

Then our integral becomes:

$$\frac{1}{m-1}\int_0^1 (1-q)^k q^\alpha dq$$

where $\alpha = \frac{1}{m-1}-1$. Now using integration by parts and induction it follows that:

$$\int_0^1 (1-q)^k q^\alpha dq = \frac{k!}{\prod_{i=1}^{k+1}\alpha+i}$$

So that, replacing α and shifting the product by one, we get:

$$p(A\,\text{never last, with alphabetical ordering}) = \frac{1}{m-1}\frac{k!}{\prod_{i=0}^{k}\frac{1}{m-1}+i}$$

Hence the probability that A is chaperoned, i.e. a last author at least once *after* the first publication is:

$$p(A\,\text{last after first paper, with alphabetical ordering}) = 1-\frac{1}{m}-\frac{1}{m-1}\frac{k!}{\prod_{i=0}^{k}\frac{1}{m-1}+i}$$

www.pnas.org/cgi/doi/10.1073/pnas.1800471115

The chaperone coefficient for a given ordering is $p(A \text{ last after first paper})$ divided by $p(A \text{ last on first paper})$. Since $p(A \text{ last on first paper})$ is the same for both orderings the ratio of the chaperone coefficients for the different orderings is just the ratio of $p(A \text{ last after first paper, with alphabetical ordering})$ divided by $p(A \text{ last after first paper, with random ordering})$, which is:

$$r(m,k) = \frac{1 - \frac{1}{m} - \frac{1}{m-1}\frac{k!}{\prod_{i=0}^{k}\frac{1}{m-1}+i}}{1 - \frac{1}{m} - (1-\frac{1}{m})^k}$$

Let us now consider this ratio for low values of k. For $k = 2$ the above ratio becomes:

$$r(m,2) = \frac{m}{2m-1}$$

As m becomes larger, this ratio rapidly approaches $r = 0.5$, already going below $r = 0.6$ for $m > 3$. We are therefore likely to observe a value of $r \approx 0.5$ in the real data, since:

- in the data from real publications the number of publications per author is likely to be heavy-tailed, with most authors having only a few publications (1);
- $k = 2$ is the minimum number of publications an author can have in order to even be considered for the chaperone effect, and given the heavy-tailed distribution this is also the most likely case;
- the average number of authors (i.e. m) for sciences and engineering is around or larger than three in the last decades (2);

Below is a table of values of r for $m = 5$ as well as the large-m limits of r, for different low k:

k	r for $m = 5$	$\lim_{m\to\infty} r$
2	0.556	0.5
3	0.498575	0.4203
4	0.466706	0.365705
5	0.448046	0.321372

We can therefore see that for a broad range of likely m and k values we would expect to see a distribution of r which peaks around 0.5 as observed in the main text Fig. 2.

2. Chaperone effect over time

In the main text, we ask whether the amount of experience acquired through chaperone bonds differs between scientific fields. Here we show data to support that the chaperone phenomenon, within each discipline, is stable over time (Figure S1), indicating that a constant amount of experience and training is required to transition between junior and senior status. Because the level of apprenticeship is fairly stable, we can meaningfully collapse the distributions to quantify differences between fields.

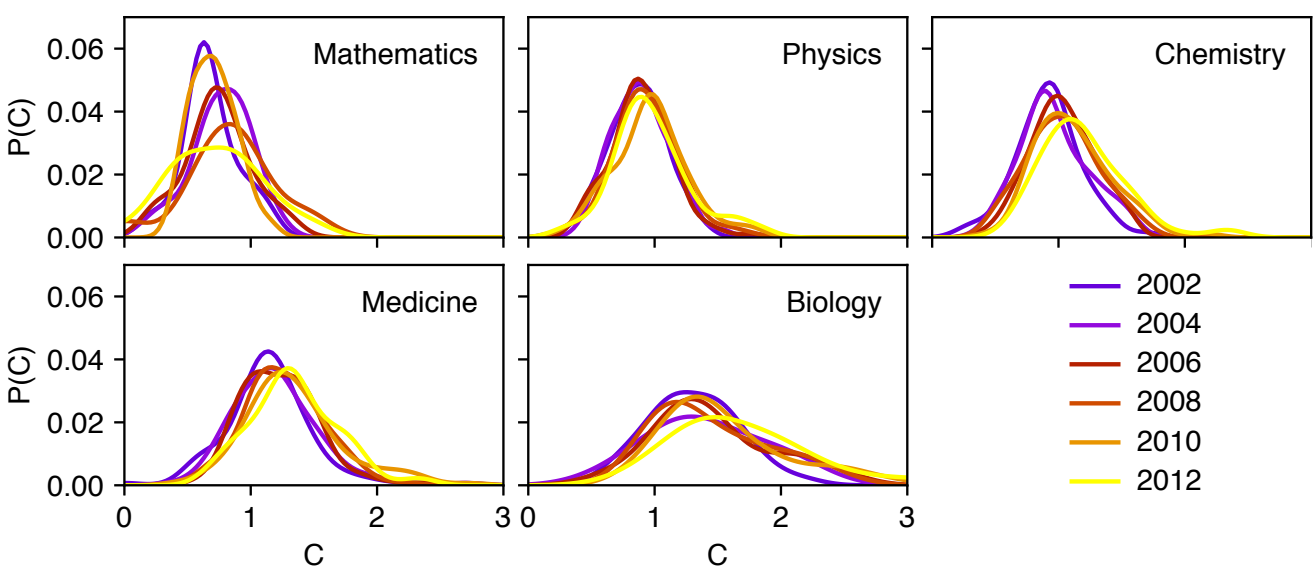

Fig. S1. Chaperone distributions for various fields as function of time. Dividing journals into branches of science reveals that the chaperone effect differs between fields and that distributions are robust across multiple years.

3. Chaperone effect across additional journals

Fig. S2 shows the key characteristics of the chaperone effect for *PNAS* and is complementary to Figs. 1-3 in the main manuscript describing the chaperone effect in *Nature*.

In the main text we also discuss the impact of *established*, *chaperoned*, and *new* authors in the case of the journal *Nature*. In Figure S3, we show the average impact ($\langle c_5 \rangle$) of publications published in *Science*, *Physical Review Letters*, *Physical Review E*, *Journal of Theoretical Biology*, *Cell*, and *American Chemical Society*.

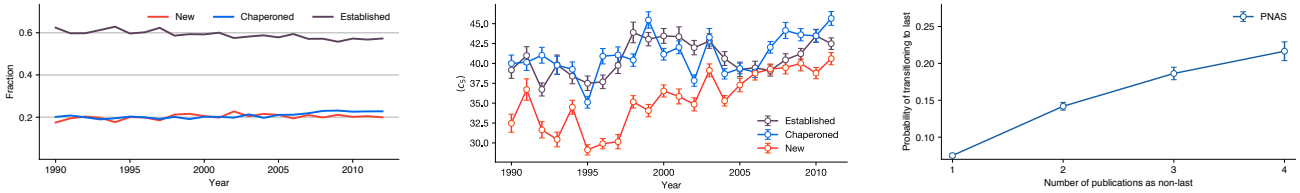

Fig. S2. (left panel) Change in author fractions over time for PNAS divided into publications with *new*, *chaperoned* and *chaperoned* last authors. (middle panel) Average impact of publications in PNAS divided into *new*, *chaperoned* and *chaperoned* last authors. Quantified using the number of citations five years after publication date ($\langle c_5 \rangle$). (right panel) The probability of transitioning to last author as function of number of occurrences as non-last author for PNAS.

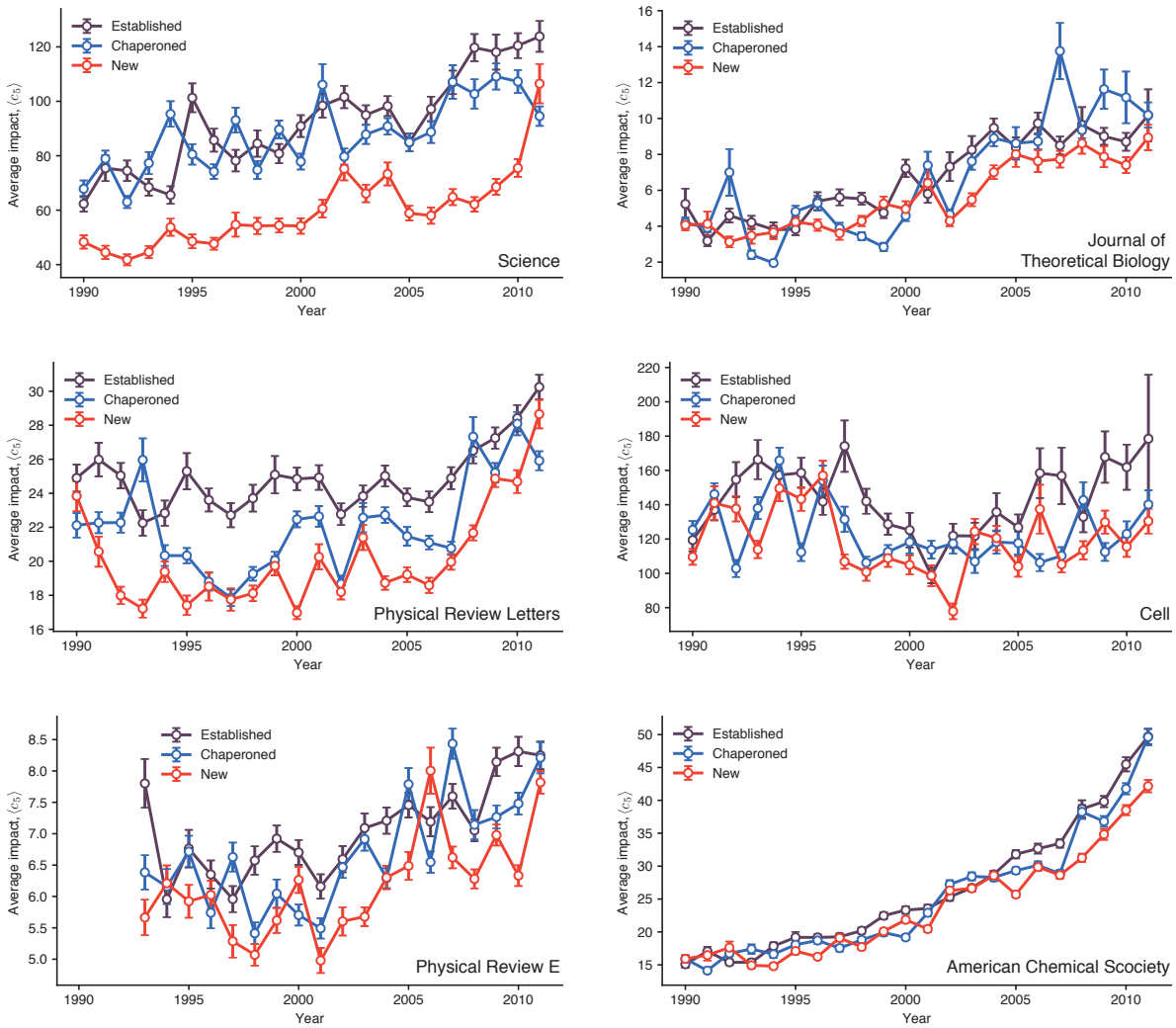

Fig. S3. Average impact ($\langle c_5 \rangle$) of publications published in *Science*, *Physical Review Letters*, *Physical Review E*, *Journal of Theoretical Biology*, *Cell*, and *American Chemical Society*. divided into new, chaperoned and established last authors.

4. Publication rates over time

Another concern is that a higher proportion of established PI-authors over time could be driven by established authors publishing more papers now than earlier. While there are some indications of increased publication rates over time for some authors (3), there is also considerable evidence that individual publication rates have remained constant over time (1, 4). To ensure, however, that our findings are not driven by increased productivity over time, we have calculated at the average total number of papers per author, per year within each journal. The results for *Nature* (which is the main panel in the Main Text Fig 1) are displayed in Figure S4.

In Figure S5, we show the number of papers published per author per year in *Nature* (note the log scale on the y-axis). It is

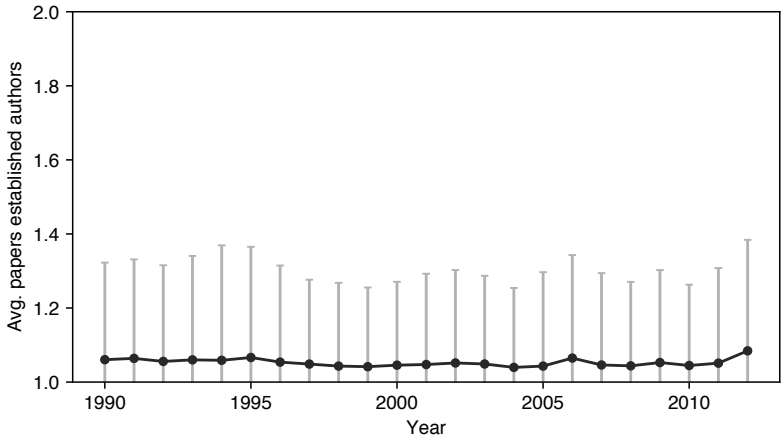

Fig. S4. Average number of papers published by established authors as PI in the journal Nature. Error bars indicate a single standard deviation.

clear from Fig. S4 that the number of papers per author in *Nature* remains roughly constant (and close to one) across the period we studied.

In order to ensure that this behavior is not unique to *Nature*, we have created similar plots for all journals, then performed linear fits to test if a similarly flat relationship exists. Fig. S5 shows the distribution of the slope of linear fits, split into fields. Overall we find no strong trends towards more papers per author within single journals – evidence that spans across fields.

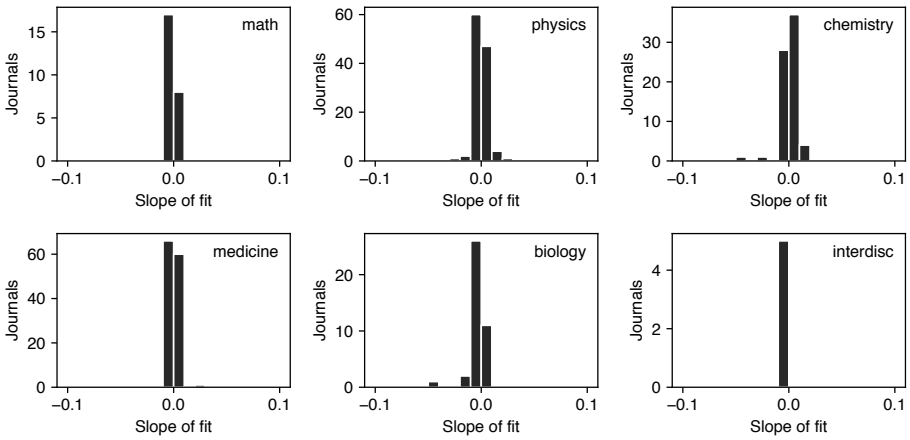

Fig. S5. Distribution of slopes after fitting a linear regression to the average number of publications for established authors per journal in each field.

5. Name disambiguation

It is crucial to have the right name disambiguation when working with citation data. In our case the problem is much less pronounced than when working with citation data more generally for two reasons: 1) we focus on authors' publications within single journals, and 2) we only focus on PIs, individuals that appear at least once as last authors on a publication. As such, some authors can appear multiple times within single journals, however, if they are never listed as a PI they will not be included in our statistics, an effect that drastically reduces the potential disambiguation noise.

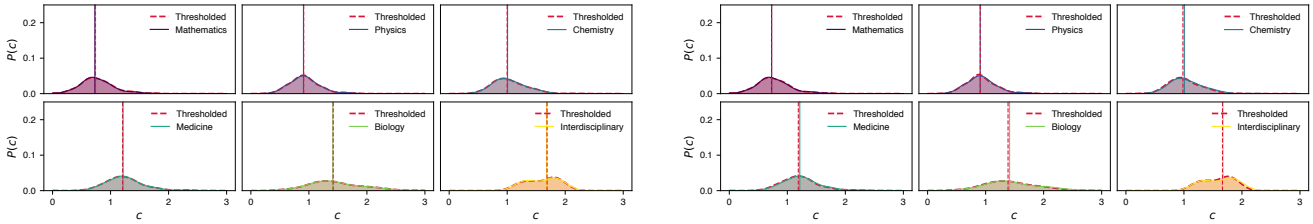

Fig. S6. Distributions of c (chaperone effect) after controlling for frequently appearing authors names. Authors are disregarded if they, on average, publish more than *two* papers per year (left panel) and *one* paper per year (right panel). Colored distributions indicate the chaperone effect taking all authors and publications into account, dashed red lines indicate c when thresholding papers with frequently appearing author names.

Take the following as example. Across all journals the top-5 most frequent names since 1990 we encounter are: "Banerjee, S.", "Brown, D. N.", "Martin, J. P.", "Alam, M. S.", and "Lee, J." of which each occurs more than 400 times (within physics journals). Yet if we look at their number of occurrences in PI roles we find that the name "Banerjee, S." appears only once in a PI role, names "Brown, D. N.", "Martin, J. P.", and "Alam, M. S." appears zero times, while "Lee, J." occurs 8 times. Thus the effect warrants further investigation.

In order to ensure that our results are not influenced by any effects caused by name-disambiguation issues, we repeated our analysis applying a threshold on author name frequency, disregarding all publications which have last authors that appear, on average, more than twice a year in a journal (this threshold is arbitrary but our analyses show that the result are not sensitive to the specific choice of threshold). This blacklists 11 770 author names across the 386 journals. We then repeated our analysis without the blacklisted author-names.

Figure S6, shows the effect on the chaperone effect distributions for each field for two different thresholds. We compare the thresholded model (dashed lines) to the original (Figure 2, main manuscript). Notice how the removal of ambiguous author names do not perceptibly change our main finding.

6. Name changes

Another potentially confounding effect is name changes due to marriage. To ensure that our findings are not driven by this effect, we study the impact of name changes on our analysis. In the literature, the topic of name switches is not addressed through analysis of publication data. One survey based study, however, limited to approx 600 female PhDs, estimated that 11.7% of the female scientists used two or more last names for their papers (5).

In order to estimate the potential impact of name-changes on our results, we have performed the following calculation for the journal *Nature*. First we unpack the publication history of all authors, then we select x percent of authors (uniformly selected) and change their names at random places in their carers. Thus, while we cannot pinpoint authors who have changed their name, we can induce name changes in our data to study its effects. Figure S7 displays the effects of changing 0, 1, 5, and 10% of all author names on the fractions of new, established, and chaperoned authors. The figure shows that authors fractions are robust even when changing 10% of author names.

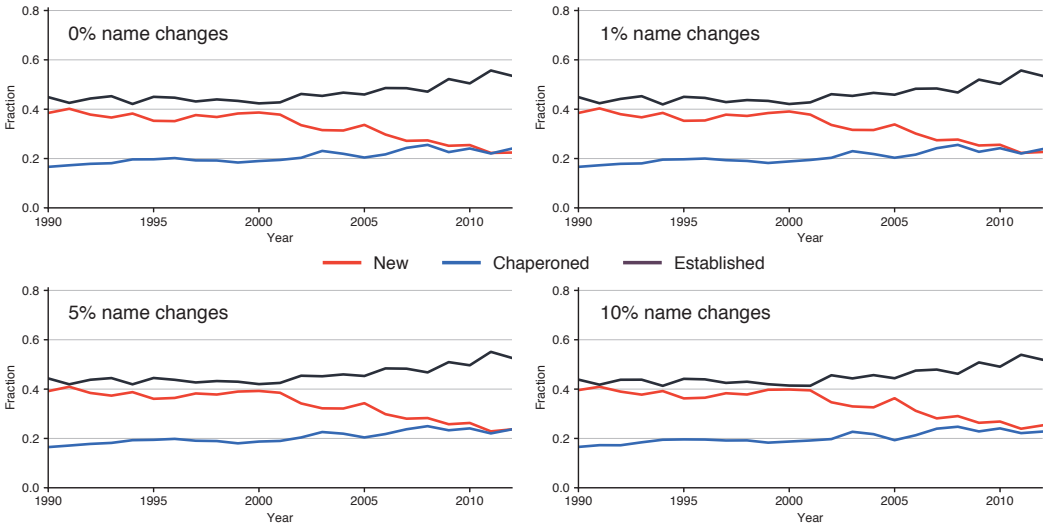

Fig. S7. Effect of name changes on new, established and chaperone authors. We show the fractions of *New*, *Established*, and *Chaperoned* for the Journal *Nature*. The fractions are the foundation of our analysis and stable wrt. changing 1, 5, and 10% of author names.

7. Alphabetical author order beyond the null model

The final issue we discuss is the fact that alphabetical ordering of authors is the norm in some fields (notable mathematics, and to some extent physics). When this is the case, the chaperone effect cannot be measured as accurately by looking at non-last → last author transitions. To better understand the importance of alphabetical ordering as the default author-order, we have repeated our analysis while removing all papers with alphabetically ordered author lists. The results are shown in Fig. S8. In this figure, the distribution of C for mathematics is smeared somewhat, but results for all other fields are similar to what we see for the full dataset.

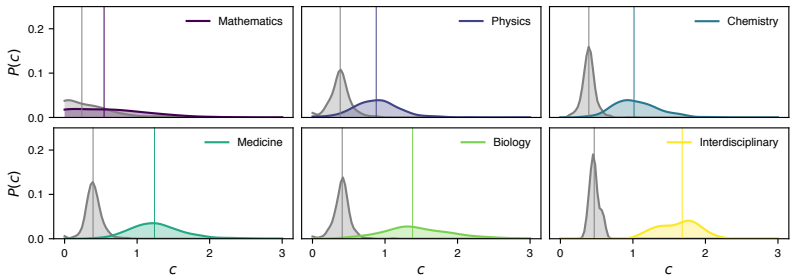

Fig. S8. Comparison of chaperone effect between scientific fields, disregarding all publications where authors are alphabetically ordered. That is, we select only papers that do not have authors in alphabetical order. Yearly distributions are then collapsed into single distributions. C is represented by the colored distributions while C_{alphabet} distributions are indicated in gray. This way we can completely separate the magnitude of the chaperone effect in the alphabetical model from the one observed in the real data.

8. SI Datasets

Data from of Web of Science cannot be shared publicly, however, we have shared aggregate statistics and code via a public GitHub repository (https://github.com/SocialComplexityLab/chaperone-open). The files and code there will allow readers to reproduce the majority of our findings. We share two types of aggregate data, 1) files containing the proportions of *new*, *established* and *chaperoned* PIs, and 2) files containing values of c, C, and C_{alphabet} over time for each journal. For both cases data is divided into individual files, one per journal, and is structured in plain text format. Data about the journal *Nature* can be downloaded for free from opensearch (https://www.nature.com/opensearch/).

We also offer the possibility to reproduce all our results starting from raw records during a research stay at Northeastern University or the Central European University.

References

1. Sinatra R, Wang D, Deville P, Song C, Barabási AL (2016) Quantifying the evolution of individual scientific impact. *Science* 354(6312):aaf5239.
2. Wuchty S, Jones BF, Uzzi B (2007) The increasing dominance of teams in production of knowledge. *Science* 316(5827):1036–1039.
3. Fanelli D, Larivière V (2016) Researchers' individual publication rate has not increased in a century. *PloS one* 11(3):e0149504.
4. Petersen AM, Jung WS, Yang JS, Stanley HE (2011) 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.
5. Long JS (1992) Measures of sex differences in scientific productivity. *Social Forces* 71(1):159–178.