Supporting Information

Petersen 10.1073/pnas.1501444112

SI Text

Aggregate Measures for Supertie Impact

In *The Apostle Effect I* and *The Apostle Effect II*, we implemented a regression model that elucidates the role of super ties at the annual level for productivity and at the paper level for citations. To provide additional quantitative evidence for the apostle effect, in this section, we develop additional descriptive measures that compare the contributions by super ties to the contributions from the rest of the collaborators.

Productivity Premium. A researcher is likely to have a relatively small number of super ties, corresponding on average to $100\langle f_R\rangle \approx 4\%$ of his/her coauthors (see Fig. 5*A*). However, these coauthors, by definition, contribute to a large fraction of the total output of i (corresponding on average to $100\langle f_N\rangle \approx$ 40–75% of all publications; see Fig. 5*B*). Thus, it is important to know the relative contributions of the super ties to nonsuper ties, because there are typically very many nonsuper tie coauthors whose inputs also contribute to the output of i.

To facilitate a comparison of productivity at the aggregate career level, we first separated the sum of the tie strengths, $K_i^T \equiv \sum_{j=1}^{S_i} K_{ij} = K_{R=1,i}^T + K_{R=0,i}^T$, into the contribution $K_{R,i}^T = \sum_{j|R_j=1} K_{ij}$ from the super ties (j with indicator value $R_j = 1$), and the complementary contribution $K_{!R,i}^T = K_i^T - K_{R=1,i}^T$ from the other ties (with $R_j = 0$). We then define the productivity premium as the ratio of the mean tie strengths,

$$p_{N,i} \equiv \frac{\langle K_{ij}|R_j=1\rangle}{\langle K_{ij}|R_j=0\rangle} = \frac{K_{R,i}^T/S_{R,i}}{K_{!R,i}^T/S_{!R,i}}, \qquad \text{[S1]}$$

between the coauthor subsets with $R_j = 1$ (totaling $S_{R,i}$ coauthors) and $R_j = 0$ (totaling $S_{!R,i} = S_i - S_{R,i}$ coauthors). This quantity increases as the ratio $S_{R,i}/S_{!R,i}$ decreases (smaller $f_{R,i}$) and as the ratio $K_{R,i}^T/K_{!R,i}^T$ increases; its maximum value is equal to the total number of publications published by the central scientist, N_i, and is bounded by the minimum value $(K_i^c + 1)/K_i^c \approx 1$ for large K_i^c.

Fig. S4*C* shows the cumulative distribution $P(\leq p_N)$. In all cases, we observe $2.5 \leq p_{N,i} \leq 33$, with average $p_{N,i}$ values between 7 and 10. Interestingly, the Top scientists from biology tend to have smaller $p_{N,i}$ values than the Other scientists (Mann–Whitney difference in median test P value = 0.0008, and K-S difference in distribution test P value = 0.0007). However, the same tests failed to indicate any significant difference for the $P(\leq p_N)$ for physics.

Citation Premium. In economic analyses, to compare nominal prices across time, it is fundamentally necessary to account for price inflation/deflation by means of an appropriate deflator index. For the same reason, it is equally important to use deflators when comparing success measures derived from other socioeconomic systems. In professional sports, for example, the rate of achievement can be era dependent—e.g., the nonstationary home run rate in Major League Baseball is an implication of the steroids era (42, 43). In science, the publication rate in physics and biology is growing at roughly a 5% rate (5). Nevertheless, this persistent growth has been subject to periods of nonstationary growth spurts, such as during the period of the US National Institutes of Health budget doubling between 1998 and 2003 (2). Thus, with these considerations in mind, in developing comparative citation measures, it is important to appropriately account for two nonstationary features of citation credit.

First, there is the time dependence of citations, arising from the fact that papers published in different years are at different points in their citation life cycle in the citation census year Y_i. The citation tallies are also affected by the underlying growth of the citation supply—due to "inflation" or "secular growth" of scientific output—which also systematically biases the comparison of raw citation counts for p from different y. Second, it is also important to divide the citation credit among the a_p coauthors of each publication p, in this way placing a cap on the net credit introduced by p, and accounting for the slow but steady exponential growth in the mean number of coauthors per paper over time (7).

To address these two underlying trends, we apply two normalizations to the raw citation count $c_{p,Y}(y)$ (measured in census year Y_i for a paper p published in year y). First, we "deflated" $c_{p,Y}(y)$ by dividing by the mean citation value for publications from the same year, $\langle c_Y^m(y)\rangle$, and then transformed this ratio into the mean citation values for the (arbitrary) baseline year $y = 2000$, giving the rescaled value

$$\tilde{c}_p \equiv c_{p,Y}(y)\frac{\langle c_Y^m(2000)\rangle}{\langle c_Y^m(y)\rangle}. \qquad \text{[S2]}$$

This also accounts for the fact that more recent publications have had less time to accrue citations than older publications. Second, we control for trends in team size, choosing a naive approach that divides the $\tilde{c}_p$ citations into equal shares among the a_p coauthors (44). As such, we define the normalized citations credited to coauthor j of p as

$$\tilde{c}_{j,p} \equiv \frac{\tilde{c}_p}{a_p}. \qquad \text{[S3]}$$

Similar to the normalization procedure used for the citation z score $z_{i,p,y}$ in Eq. **7**, $\langle c_Y^m(y)\rangle$ is the average number of citations for publications published in a benchmark set m, choosing m to be the aggregation of articles appearing in the multidisciplinary journals *Nature*, *Proceedings of the National Academy of Sciences*, and *Science*. We restricted our query to publications denoted as "Articles," which excludes reviews, letters to the editor, corrections, and other content types. We use these high-impact journals because they have high citation rates and hence provide a robust detrending baseline for the time-dependent component of $c_{p,Y}(y)$. Again, the choice of baseline year $y = 2000$ is arbitrary (as is the deflation year 2000 commonly used in economic analyses) and is mainly used to recover the units of citations for the $\tilde{c}$ measure. Because the constant factor $\langle c_Y^m(2000)\rangle$ is used for all $\tilde{c}_p$ values, it does not affect our results. The advantage of $\tilde{c}_p$ over $z_{i,p,y}$ is that the former is a positive number including the value 0, and hence can be added across p; $z_{i,p,y}$, however, can be negative and is centered around 0, and, therefore, summing across p has a different interpretation that is not suitable for what follows.

We define the cumulative measure of citation impact for coauthors i and j as

$$\tilde{C}_{i,j} \equiv \sum_{p \text{ with } j} \tilde{c}_{j,p}, \qquad \text{[S4]}$$

where the sum includes only those publications in the profile of i that also include coauthor j. In the extreme case that j is a coauthor of every publication, $K_{ij} = N_i$, this pairwise measure has

the upper limit equal to the citation share of the central scientist, $\tilde{C}_{i,i} \equiv \sum_p \tilde{c}_{i,p}$. The sum across all j including i, $\tilde{C}_i = \tilde{C}_{i,i} + \sum_j \tilde{C}_{i,j}$, yields the net detrended citation value, which is independent of the distribution of a_p.

To define a similar citation premium, we also separated the citations into the contributions from the $S_{R,i}$ super ties and the contributions from the $S_{!R,i}$ nonsuper tie collaborators. Because the total $\tilde{C}_i$ is conserved, we split the $\tilde{C}_{i,j}$ into two groups: The total for the coauthors with $R_j = 1$ is $\tilde{C}_{R,i} \equiv \sum_{j|R=1} \tilde{C}_{i,j}$, and the total for the remaining coauthors is $\tilde{C}_{!R,i} \equiv \sum_{j|R=0} \tilde{C}_{i,j}$. We then define the citation premium to be the ratio of the average citation shares of the coauthors in each subset,

$$p_{C,i} \equiv \frac{\langle \tilde{C}_{R,i} \rangle}{\langle \tilde{C}_{!R,i} \rangle} = \frac{\tilde{C}_{R,i}/S_{R,i}}{\tilde{C}_{!R,i}/S_{!R,i}}, \qquad \text{[S5]}$$

which has a minimum possible value equal to 0 and, in principle, has no upper bound. Fig. S4*D* shows the distribution of $p_{C,i}$, with mean, median, and maximum values across all datasets of 14.1, 11.3, and 134, respectively. We observed only two profiles (2 out of 473) with $p_{C,i} < 1$. Thus, using a group-to-group comparison, this measure shows that the relative citation impact contribution of super ties to other ties is significantly greater than unity. There may be a self-selection, because high-quality work may induce follow-up research, presumably with a similar set of collaborators. Hence, the citation premium is also evidence for the value of persistent collaboration, which can leverage and build upon prior experience and cumulative pairwise achievement.

Also of interest, we observe a consistent pattern considering the distributions of both $p_{N,i}$ and $p_{C,i}$: The Top scientist profiles have smaller mean values than their counterparts, and the biology profiles have smaller mean value than for physics. In the case of productivity, this may follow from their privileged access to short-term collaboration opportunities. In the case of the citation impact, this pattern may emerge due to the reputation asymmetry of top scientists, who, by way of their prestige, may have more control over their choice of collaborators, possibly aimed at reducing redundancy within the team, reducing the team size, which also increases the citation credit per coauthor, $\tilde{c}_{j,p}$. In large-team efforts, because most collaboration durations are short with relatively small K_{ij}, increasing a_p is most likely to decrease $p_{C,i}$ by way of decreasing the numerator and increasing the denominator.

Because $p_{C,i}$ is an aggregate career measure, and the dependent variable $z_{i,p,y}$ in our citation regression model (Eq. **7**) is a normalized measure that does not have the dimensionality of citations, it is difficult to use these quantities to measure the citation boost on a per-publication basis. Thus, to estimate the apostle effect on the long-term citation tally of individual publications, we separated the set of publications with at least one super tie coauthor ($R_p = 1$) from the complementary set of publications without any super tie coauthors ($R_p = 0$). To compare p from a similar era, we took all of the publications from the 11-y window 1990–2000. Also, because citation rates are discipline dependent, we distinguished between biology and physics publications. During this period, 62% (7,814) of the p have $R_p = 1$ for biology and 57% (10,128) of the p have $R_p = 1$ for physics. From these well-balanced subsets, we then estimated the citation impact due to $R_p = 1$ in two ways.

First, we calculated the cumulative citation distribution, $P(\tilde{c}|R_p)$, for the publications with $R_p = 0,1$. Fig. S6 *A* and *B* shows each distribution on log-linear axes, which emphasizes the log-normal features of $P(\tilde{c})$. On this log-linear scale, the two distributions are characterized by a horizontal offset, which is visible for the majority of the $\tilde{c}_p$ range. This graphical feature indicates that, in distribution, the $\tilde{c}_p$ for $R_p = 1$ are larger by an approximately constant factor α_R, i.e., $P(\tilde{c}|R_p = 1) \approx P(\alpha_R \tilde{c}|R_p = 0)$. We estimate α_R by comparing the means and the median values of the $P(\tilde{c})$ distributions. For example, the ratio between the means yields the value $\alpha_R = \langle \tilde{c}_p | R = 1 \rangle / \langle \tilde{c}_p | R = 0 \rangle = 1.17$ for biology and 1.16 for physics. Estimating α_R using the ratio of the median values yields approximately the same value. Thus, α_R represents a 16–17% citation boost for p with $R_p = 1$. For the average-cited p, this boost translates to a 21-citation difference for biology and an 8-citation difference for physics. These numbers, however, arise from an aggregated dataset, so it is not necessarily true that α_R is representative of all scientists.

To confirm the per-publication citation premium at the researcher level, we grouped the publications with $R_p = 0,1$ within each profile i. To reduce the sensitivity to fluctuations, we analyzed only the i with at least 10 publications in the $R_p = 0$ subset and at least 10 publications in the $R_p = 1$ subset. Then, to obtain a characteristic citation measure for each the two $R_p = 0,1$ subsets, we calculated the median value, $\tilde{c}_{R,i}$, for the subset of p with $R_p = 1$, and the median value, $\tilde{c}_{!R,i}$, for the complementary publication subset with $R_p = 0$.

Fig. S6 *C* and *D* shows the scatter plot of $\tilde{c}_{!R,i}$ and $\tilde{c}_{R,i}$ for each i. The line $y = x$ distinguishes the researchers with $\tilde{c}_{R,i} > \tilde{c}_{!R,i}$. There is notable heterogeneity across the i in terms of the citation premium from super ties. Nevertheless, the majority of researchers have $\tilde{c}_{R,i} > \tilde{c}_{!R,i}$, with 73% of the biology researchers and 76% of the physics researchers above the $y = x$ line. We then obtained a second estimate of the per-publication citation premium by fitting a least-squares model, $\tilde{c}_{R,i} = \mu \tilde{c}_{!R,i} + \epsilon$, where ϵ is an ordinary least squares (OLS) error term, obtaining best-fit values $\mu = 1.21 \pm 0.06$ (biology) and $\mu = 1.24 \pm 0.09$ (physics).

Thus, these last two methods provide consistent estimates of the citation boost at the publication level, $\mu \approx \alpha_R$ corresponding to a 16–24% citation boost, pointing to a significant long-term citation impact attributable to the presence of super ties.

Data Description

Name Disambiguation Strategy. We obtained the top-cited researcher publication data using the Distinct Author Sets function provided by TRWOK to increase the likelihood that only publications actually authored by each central author i are analyzed. On a case by case basis, we performed further author disambiguation within each profile. The Other (matched set) profiles were also downloaded from TRWOK, either by using the Distinct Author database option, or by collecting distinct researcher profile data from ResearcherID.com.

In this latter case of ResearcherID.com profiles, we collected biology and physics profiles by querying the database for profiles listing any of the following keywords: graphene, neuroscience, molecular biology, or genomics. For further details on the selection procedure and for extensive analysis of the statistical properties of these datasets, see the data descriptions in refs. 45–47.

The data census year Y_i refers to the calendar year in which the researcher profile data were downloaded. Let y_i^0 be the first calendar year of his/her first publication and y_i^f be the calendar year of the last observed publication, so that the total number of years of data for i is $T_i = y_i^f - y_i^0 + 1$. Hence, depending on if the career i was completed in Y_i, there are two possible scenarios relating Y_i and T_i: (scenario a) if the researcher i was still active in Y_i, then $Y_i = y_i^f = y_i^0 + T_i - 1$ and $\Delta Y_i = Y_i - y_i^f = 0$; or (scenario b) if his/her career terminated at some time before Y_i, then $Y_i = y_i^0 + T_i - 1 + \Delta Y_i$, with $\Delta Y_i > 0$ and T_i corresponding to the final career length. The datasets comprise profiles with census year Y_i varying from 2010 to 2012 (47). These relatively small variations in Y_i do not alter the citation results because all citation measures are appropriately detrended to make possible comparisons across time. Moreover, the regression data are longitudinal, meaning that the observations are made according to t, and so the results do not depend on T_i or the completeness

of the career. Furthermore, the regression models each include an author-level fixed-effect parameter $\beta_{i,0}$ that controls for time-invariant author-specific properties, thereby absorbing factors related to the starting calendar year y_i^0 and the lag ΔY_i.

For a given central author i, we aggregate the TRWOK publications and create a registry of surname and first/middle-initial pairs, {Surname, FM}, where FM can consist of one, two, or three alphabetic first-letter character abbreviations α, FM $\equiv \alpha_1\alpha_2\alpha_3$. Because the number of distinct coauthors per i is relatively small, on the order of 10–1,000 distinct names per profile, we assume that a name disambiguation problem among the coauthors does not introduce significant levels of type 1 "splitting" or type 2 "clumping" disambiguation errors. Hence, we perform a string matching on similar last names and α_1, ignoring α_2 and α_3 so that publications with variable listing of α_2 and α_3 do not result in a type 1 "profile splitting" error. We then aggregate the publication information into the profile of coauthor j of central author i. Because our approach is egocentric, we do not analyze the publications of j that do not include i. Clearly, this would require nearly comprehensive TRWOK publication data, which is a major data limitation.

Matched Profile Selection Criteria. To account for possible prestige effects, we compared top-cited profiles to a set of Other profiles that we matched within each discipline. To match the datasets, we collected "not top-cited" researcher profiles that had levels of career length and productivity similar to the top-cited profiles. More specifically, we introduced a productivity criteria requiring that an Other profile must have at least as many publications, N_i, as all of the researchers in the corresponding top-cited dataset: For biology, this minimum threshold value is $Min(N_i|\text{top}-\text{cited}) = 52$, and for physics, it is $Min(N_i|\text{top}-\text{cited}) = 46$. Altogether, our career dataset comprised 100 top-cited and 93 matched profiles from biology, and 100 top-cited and 180 matched profiles from physics.

Throughout our analysis, we introduced various quantities that summarize the career (career length T_i, total publications N_i, etc.) and collaboration pattern (mean duration $\langle L_i \rangle$, mean strength $\langle K_i \rangle$, strength Gini coefficient G_i, etc.) of any given research profile i. We found that the Top and Other datasets are statistically well-matched with respect to some variables, using the K-S test to certify the null hypothesis that the underlying distributions are statistically similar. For example, the super tie coauthor fraction $f_{R,i}$ exhibits the same distribution across all four datasets, as shown in Fig. 5*A*. Other variables were well matched only within discipline, e.g., $\langle K_i \rangle$, or were well matched only within Top or Other datasets, e.g., $f_{K,i}$.

One variable worth mentioning, for which the Top and Other datasets were not well matched, was the career length distribution, $P(T_i)$. Because the Top scientists were selected on account of cumulative citation tallies, they are biased toward longer T_i, many of which are completed careers. Because the maximum possible L_{ij} is given by T_i, the $\langle L_i \rangle$ variables may be biased toward longer values for the top-cited researcher profiles. As such, we avoid making any comparisons on account of this type of measure. Instead, our comparisons in the manuscript are based on more intensive measures, e.g., the super tie coauthor fraction f_R, which are less sensitive to biases arising from systematic differences in T_i and ΔY_i.

Moreover, our analysis of the apostle effect, by design, avoids the potential bias due to T_i. For example, the productivity premium $p_{N,i}$ and the citation premium $p_{N,i}$ are ratios in which both the numerator and the denominator should have approximately the same dependence on T_i, and so the effect cancels out. In the case of the regression models, the dependent and independent variables are all specific to a particular career year t.

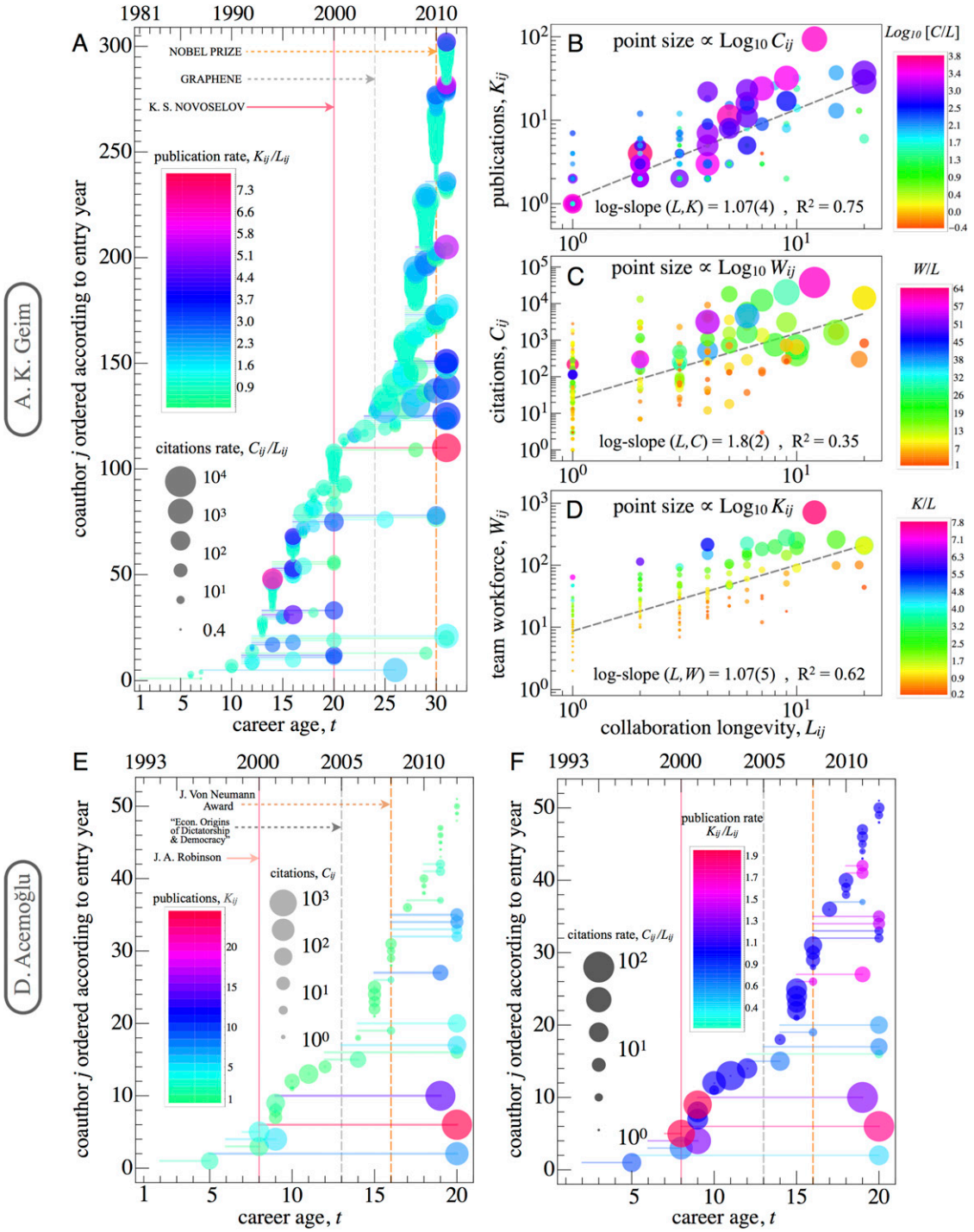

Fig. S1. Complex relations between productivity, collaboration, and impact. *A–D* are for A. K. Geim, who is characterized by an average collaboration duration of 2.1 y (calculated including the collaborations with $L_{ij} = 1$ but excluding the collaborations active in the last 2 y), a characteristic tie strength $\langle K_i \rangle = 3.7$ publications, a collaboration radius of $S_i = 303$ coauthors, and $N_i(2012) = 217$ total publications; *E–F* are for D. Acemoglu, who is characterized by an average collaboration duration of 1.6 y (also calculated including the collaborations with $L_{ij} = 1$ but excluding the collaborations active in the last 2 y), $\langle K_i \rangle = 2.9$ publications, $S_i = 51$ coauthors, and $N_i(2012) = 118$ publications. These schematics demonstrate how the visualization of dynamic ego network changes if we use publication and citation measures that are normalized by L_{ij}, resulting in per-year-of-collaboration (intensity) measures. (*A*) Collaboration measures calculated per unit time, for comparison with Fig. 1. (*B–D*) Scatter plots for the profile of A. K. Geim relating collaboration duration (L_{ij}), with (*B*) collaboration strength (K_{ij}), (*D*) pairwise team size (W_{ij}), and (*C*) citations (C_{ij}). W_{ij} is the total number of coauthors (nondistinct) on publications including *i* and *j*, a proxy for pairwise collaborative input, conditioned on *i* and *j*. The dashed line in each panel represents the ordinary least-squares fit of the log of the variables. As such, the logarithmic slope (scaling exponent) is listed in each panel, and the value in parentheses represents the SE in the last digit reported. (*E* and *F*) Economics is a field not traditionally considered to be collaborative at the rates of physics or biology. Nevertheless, prestige and collaboration life cycles are still important factors, independent of discipline. To demonstrate this, we show the career profile of the highly cited economist, Daron Acemoglu. Notable landmark achievements are indicated, including the early partnership with James A. Robinson in 2000, and their groundbreaking book, *Economic Origins of Dictatorship and Democracy*, published in 2005 (48). (*E*) Net collaboration measures for D. Acemoglu, analogous to Fig. 1. (*F*) Collaboration measures calculated per unit time, analogous to *A*.

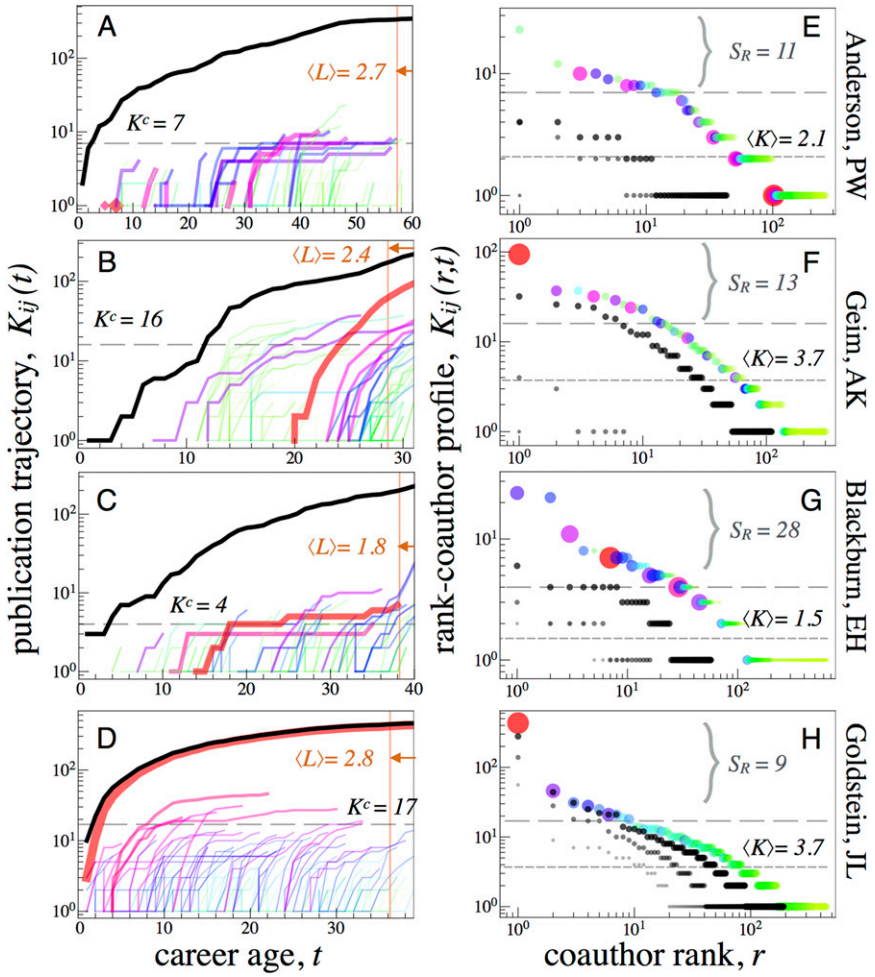

Fig. S2. Visualizing the dynamic collaboration profile of individual researchers: the longitudinal coauthor trajectories of (*A*) Anderson, (*B*) Geim, (*C*) Blackburn, and (*D*) Goldstein; the cross-sectional rank-citation profiles of (*E*) Anderson, (*F*) Geim, (*G*) Blackburn, and (*H*) Goldstein. For each discipline, we show the collaboration profile of two Nobel laureates (A. K. Geim and J. L. Goldstein) whose top-cited research was done with their most intense collaborator, and two collaboration profiles for two Nobel laureates (P. W. Anderson and E. H. Blackburn) whose top-cited research did not exhibit this feature. Despite their common achievement, we observe a wide variation in the entry, strength, and saturation of their collaborations. To illustrate the variation in tie strength, both within and between researcher profiles, we show the rank–coauthor profile $K_{ij}(r,t)$, which is defined for any given t by sorting the coauthors in decreasing order by rank r, $K_{ij}(r=1,t) \geq K_{ij}(2,t) \geq \ldots \geq K_{ij}(r=S_i,t)$. In this way, $K_{ij}(r,t)$ provides a cross-sectional representation of $K_{ij}(t)$. As such, snapshots of $K_{ij}(r,t)$ taken at different t capture the temporal evolution of a researcher's tie strength distribution, as illustrated by the gray data points in *E*–*H*. (*A*–*D*) Longitudinal growth of $K_{ij}(t)$, the cumulative number of publications with coauthor j (colored curves), and the central author's total number of publications $N_i(t)$ (black curve). To reduce graphical clutter, we truncate each $K_{ij}(t)$ at the year of the last observed collaboration; otherwise, each panel would be dominated by horizontal lines. The gray dashed line indicates K_i^c, which distinguishes the $K_{ij}(t)$ trajectories corresponding to super ties. The distance between the vertical yellow line and the right edge of each panel indicates the mean collaboration duration, $\langle L_i \rangle$, for each researcher. (*E*–*H*) To convey the dynamics of the rank-coauthor profile, we show snapshots of $K_{ij}(r,t)$ for $t = 5$ y, 10 y, and 20 y (increasing gray dot size), in addition to the final $K_{ij}(r, t = T_i)$ (colored circles) calculated for the most recently available career year T_i. The lower dashed gray line indicates $\langle K_i \rangle$, which separates the weak from the strong ties. The upper dashed gray line indicates K_i^c, which distinguishes the $S_{R,i}$ super ties within the subset of strong ties. Recently, the analog of the *h*-index has been suggested as a way to measure the "author core" derived from the rank-coauthor distribution (49). For all panels, to facilitate visual comparison, the color scale used in the left and right column is the same for each i. To identify the coauthors with the highest net citation impact, we plot curves (circles) using thickness (radius) and color that are scaled proportional to $\log \bar{C}_{i,j}$, which is the log of the total citation share of coauthor j in profile i (see Eq. **S4**).

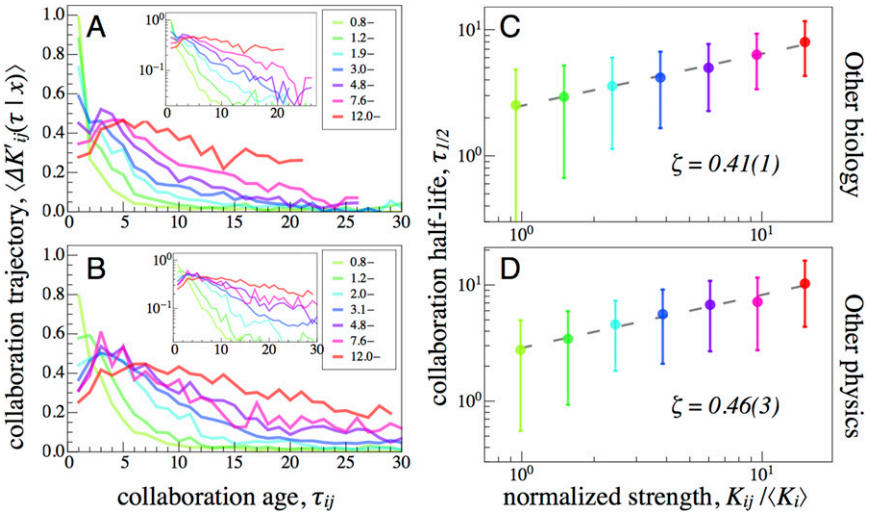

Fig. S3. Collaboration life cycle for the (*A* and *C*) Other biology and the (*B* and *D*) Other physics datasets. Other datasets: (*A* and *B*) Average collaboration strength, normalized to peak value, measured τ years after the initiation of the collaboration tie. (*Insets*) On log-linear axes, the decay appears as linear, corresponding to an exponential form. (*C* and *D*) For each $\{x\}$ group, we show the average and SD (error bar) of $\tau_{1/2}$; we use logarithmically spaced $\{x\}$ groups that correspond by color to the same $\{x\}$ as in *A* and *B*. The ζ value quantifies the scaling of $\langle \tau_{1/2} \rangle$ as a function of the normalized coauthor strength $x \equiv K_{ij}/\langle K_i \rangle$. The sublinear ($\zeta < 1$) values indicate that collaborations are distributed over a timescale that grows slower than proportional to x; conversely, this means that longer collaborations are more productive, being characterized by increasing marginal returns ($1/\zeta > 1$). Fig. 3 shows the analogous plot for the Top physics and biology datasets; all four datasets exhibit similar features.

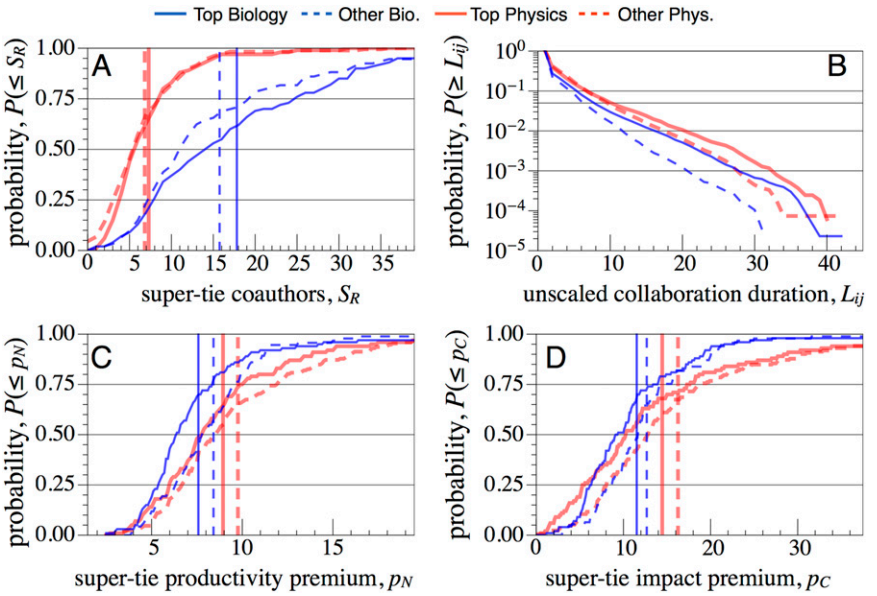

Fig. S4. Additional collaboration profile measures. (*A*) Cumulative distribution of the number of super ties $S_{R,i}$. The mean (vertical lines) and SD are 18 ± 13 (Top biology), 16 ± 13 (Other biology), 7.3 ± 4.8 (Top physics), and 6.8 ± 5.1 (Other physics). The K-S test *P* value calculated by comparing the biology distributions is 0.12, and, for the physics distributions, it is 0.34; in both cases, the null hypothesis that the two compared datasets arise from the same distribution is not rejected at the 5% level. (*B*) Cumulative distribution of the empirical (unnormalized) durations L_{ij} (years). The $L_{ij} = 1$ values dominate the distribution, with $P(L_{ij} = 1) =$ 0.73 y (Top biology), 0.78 y (Other biology), 0.61 y (Top physics), and 0.58 y (Other physics). Thus, including the $L_{ij} = 1$ values, the mean L_{ij} are 2.2 y (Top biology), 1.8 y (Other biology), 2.7 y (Top physics), and 2.7 y (Other physics). To avoid age cohort bias, collaborations commenced in the final L_i^c period of each career profile are excluded from these distributions. (*C*) Cumulative distribution of the productivity premium p_N defined in Eq. **S1**. The mean and SD are 7.6 ± 4.4 (Top biology), 8.4 ± 3.6 (Other biology), 8.9 ± 4.8 (Top physics), and 9.8 ± 4.5 (Other physics). Only the two physics datasets are significantly similar (K-S $p = 0.35$). (*D*) Cumulative distribution of the citation premium p_C defined in Eq. **S5**. The mean and SD are: 12 ± 10 (Top biology), 13 ± 7 (Other biology), 15 ± 16 (Top physics), and 16 ± 14 (Other physics). The K-S test *P* values calculated by comparing the two Top datasets and the two Other datasets are both greater than 0.05. An interesting and consistent pattern emerges when considering the distributions of both p_N and p_C: The Top scientist profiles have smaller mean values than their counterparts, and the biology profiles have smaller mean value than for physics. The mean, median, and maximum values across all datasets are 14.1, 11.3, and 134, respectively, with all but two values greater than unity. Because the maximum value is an extreme outlier, we truncate the *x* axes showing only values of <38, which represents more than 95% of the data.

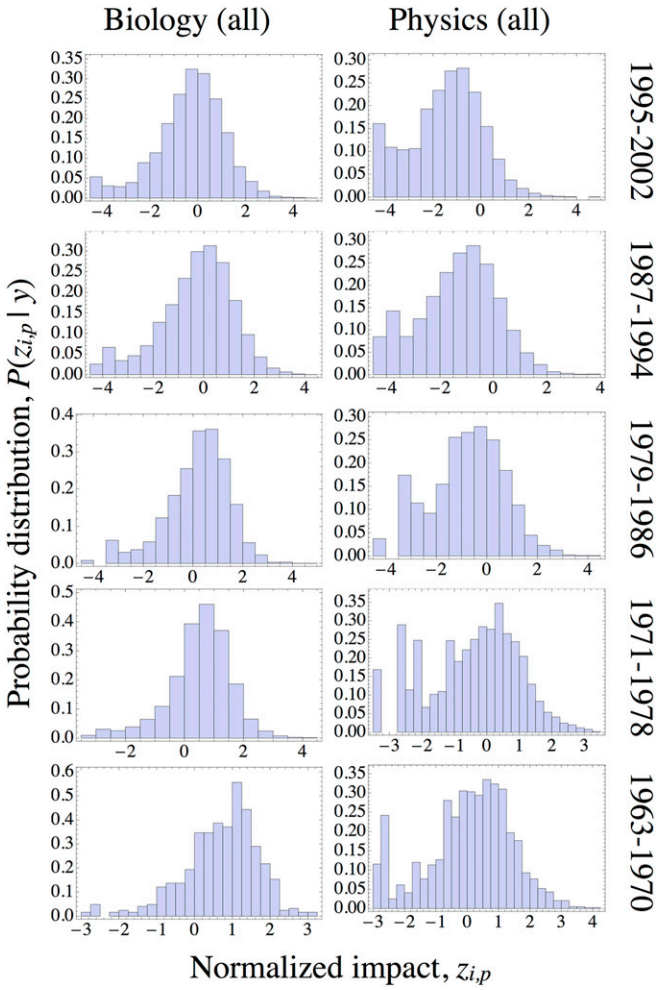

Fig. S5. Distribution of normalized citation impact z. Each panel shows the pdf $P(z|y)$ using z values aggregated over successive nonoverlapping 8-y periods. These panels demonstrate the distribution stability of $P(z|y)$ over time, where z is the dependent variable in the citation apostle effect model in Eq. **8**.

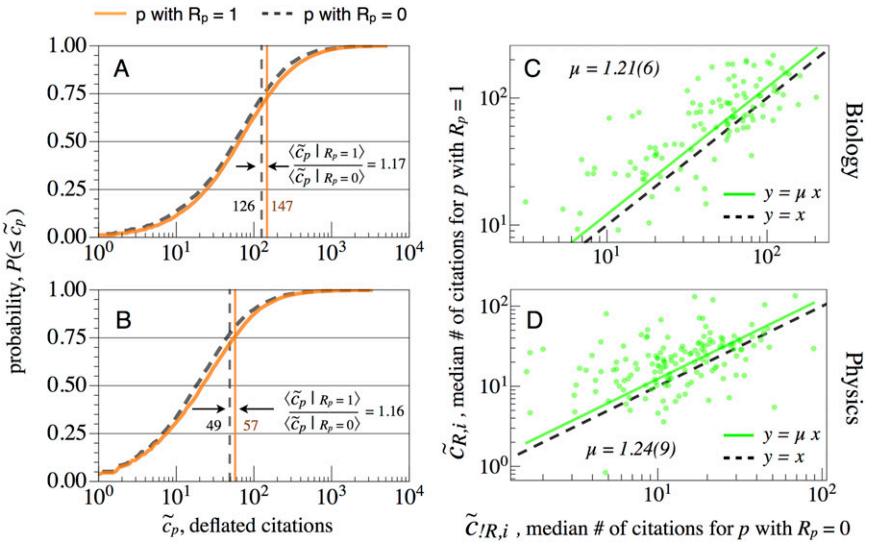

Fig. S6. Comparing the citation distribution for papers with and without super ties: (*A* and *C*) Top and Other biology datasets combined, and (*B* and *D*) Top and Other physics datasets combined. (*A* and *B*) The cumulative citation distribution, $P(\tilde{c})$, of the detrended citations $\tilde{c}_p$ defined in Eq. **S2**. The solid orange curve represents $P(\tilde{c})$ for publications with $R_p = 1$, and the dashed black curve represents $P(\tilde{c})$ for publications with $R_p = 0$. Pairwise comparison of the distributions yield K-S *P* values less than 10^{-6}, indicating that the distributions are significantly different. The distribution means are indicated by the vertical lines with corresponding numerical value shown in each panel. The ratio between the means yields the value $\alpha_R = \langle \tilde{c}_p | R = 1 \rangle / \langle \tilde{c}_p | R = 0 \rangle = 1.17$ for biology and 1.16 for physics. Estimating α_R using the ratio of the median values yields approximately the same value. Thus, α_R represents a 16–17% citation boost for *p* with $R_p = 1$, which translates, on average, to a 21-citation difference for biology and an 8-citation difference for physics. (*C* and *D*) Scatter plots of the median $\tilde{c}_{p,i}$ values for *p* with $R_p = 1$ versus the median $\tilde{c}_{p,i}$ values for *p* with $R_p = 0$. Values are calculated within researcher profiles; thus each dot represents a single researcher. The majority of researchers have $\tilde{c}_{R,i} > \tilde{c}_{!R,i}$, with 73% of the biology researchers and 76% of the physics researchers above the (dashed black) $y = x$ line. The μ value estimates the per-publication citation premium that accounts for heterogeneity across *i*. Because $\alpha_R \approx \mu$, these two methods yield consistent estimates of the citation premium per publication.

Table S1. Apostle effect productivity model ($n_{i,t}$): Parameter estimates for the fixed-effects regression model in Eq. 6 with $\Delta t = 3$-y-long periods, using robust SEs implemented by the Huber/White/sandwich method

Dataset	*A*	$\ln \bar{a}_t$	$\bar{L}_t$	G_t^K	ρ_t	*t*	$N_{obs.}$	Adj. R^2
All	406	**0.127 ± 0.044**	**−0.078 ± 0.013**	**1.060 ± 0.125**	**0.152 ± 0.026**	**0.029 ± 0.003**	2,890	0.16
(Std. coeff.)		**0.169 ± 0.059**	**−0.218 ± 0.038**	**0.268 ± 0.032**	**0.176 ± 0.030**	**0.060 ± 0.005**		
P value		0.004	0.000	0.000	0.000	0.000		
Biology (Top)	99	−0.149 ± 0.092	−0.059 ± 0.045	**3.003 ± 0.406**	**0.175 ± 0.071**	**0.035 ± 0.005**	782	0.24
P value		0.110	0.199	0.000	0.016	0.000		
Biology (Other)	84	0.126 ± 0.094	−0.067 ± 0.041	**2.159 ± 0.504**	0.080 ± 0.055	**0.047 ± 0.008**	492	0.31
P value		0.184	0.104	0.000	0.146	0.000		
Physics (Top)	99	−0.073 ± 0.112	**−0.086 ± 0.022**	**1.918 ± 0.426**	**0.159 ± 0.036**	**0.024 ± 0.004**	753	0.11
P value		0.514	0.000	0.000	0.000	0.000		
Physics (Other)	124	0.152 ± 0.076	**−0.072 ± 0.022**	**1.514 ± 0.327**	**0.160 ± 0.043**	**0.025 ± 0.006**	863	0.13
P value		0.047	0.001	0.000	0.000	0.000		

See Table 2 for results with $\Delta t = 1$. Only profiles with four or more data values were included in the regression. Values significant at the $p < 0.02$ level are indicated in boldface.

Other Supporting Information Files

Dataset S1 (XLSX)