Quantifying the impact of weak, strong, and super ties in scientific careers

Alexander Michael Petersen[1]

Laboratory for the Analysis of Complex Economic Systems, Institutions, Markets, Technologies (IMT) Lucca Institute for Advanced Studies, 55100 Lucca, Italy

Edited by Yu Xie, University of Michigan, Ann Arbor, MI, and approved July 1, 2015 (received for review January 22, 2015)

Scientists are frequently faced with the important decision to start or terminate a creative partnership. This process can be influenced by strategic motivations, as early career researchers are pursuers, whereas senior researchers are typically attractors, of new collaborative opportunities. Focusing on the longitudinal aspects of scientific collaboration, we analyzed 473 collaboration profiles using an egocentric perspective that accounts for researcher-specific characteristics and provides insight into a range of topics, from career achievement and sustainability to team dynamics and efficiency. From more than 166,000 collaboration records, we quantify the frequency distributions of collaboration duration and tie strength, showing that collaboration networks are dominated by weak ties characterized by high turnover rates. We use analytic extreme value thresholds to identify a new class of indispensable super ties, the strongest of which commonly exhibit >50% publication overlap with the central scientist. The prevalence of super ties suggests that they arise from career strategies based upon cost, risk, and reward sharing and complementary skill matching. We then use a combination of descriptive and panel regression methods to compare the subset of publications coauthored with a super tie to the subset without one, controlling for pertinent features such as career age, prestige, team size, and prior group experience. We find that super ties contribute to above-average productivity and a 17% citation increase per publication, thus identifying these partnerships—the analog of life partners—as a major factor in science career development.

computational social science | cooperation | team science | career evaluation | bibliometrics

Science operates at multiple scales, ranging from the global and institutional scale down to the level of groups and individuals (1). Integrating this system are multiscale social networks that are ripe with structural, social, economic, and behavioral complexity (2). A subset of this multiplex is the scientific collaboration network, which forms the structural foundation for social capital investment, knowledge diffusion, reputation signaling, and important mentoring relations (3–8).

Here we focus on collaborative endeavors that result in scientific publication, a process that draws on various aspects of social ties, e.g., colocation, disciplinary identity, competition, mentoring, and knowledge flow (9). The dichotomy between strong and weak ties is a longstanding point of research (10). However, in "science of science" research, most studies have analyzed macroscopic collaboration networks aggregated across time, discipline, and individuals (11–21). Hence, despite these significant efforts, we know little about how properties of the local social network affect scientists' strategic career decisions. For example, how might creative opportunities in the local collaboration network impact a researcher's decision to explore new avenues versus exploiting old partnerships, and what may be the career tradeoffs in the short versus the long term, especially considering that academia is driven by dynamic knowledge frontiers (22, 23).

Against this background, we develop a quantitative approach for improving our understanding of the role of weak and strong ties, meanwhile uncovering a third classification—the super tie—which we find to occur rather frequently. We analyzed longitudinal career data for researchers from cell biology and physics, together comprising a set of 473 researcher profiles spanning more than 15,000 career years, 94,000 publications, and 166,000 collaborators. To account for prestige effects, we define two groups within each discipline set, facilitating a comparison of top-cited scientists with scientists who are more representative of the entire researcher population (henceforth referred to as "Other"). From the N_i publication records spanning the first T_i career years of each central scientists i, we constructed longitudinal representations of each scientist's coauthorship history.

We adopt an egocentric perspective to track research careers from their inception along their longitudinal growth trajectory. By using a local perspective, we control for the heterogeneity in collaboration patterns that exists both between and within disciplines. We also control for other career-specific collaboration and productivity differences that would otherwise be averaged out by aggregate cross-sectional methods. Thus, by simultaneously leveraging multiple features of the data—resolved over the dimensions of time, individuals, productivity, and citation impact—our analysis contributes to the literature on science careers as well as team activities characterized by dynamic entry and exit of human, social, and creative capital. Given that collaborations in business, industry, and academia are increasingly operationalized via team structures, our findings provide relevant quantitative insights into the mechanisms of team formation (15), efficiency (24), and performance (25, 26).

The organization of our study is structured as follows. The longitudinal nature of a career requires that we start by quantifying the tie strength between two collaborators from two different perspectives: duration and strength. First we analyze the collaboration duration, L_{ij}, defined as the time period between the first and last publication between two researchers i and j. Our

Significance

A scientist will encounter many potential collaborators throughout his/her career. As such, the choice to start or terminate a collaboration can be an important strategic consideration with long-term implications. While previous studies have focused primarily on aggregate cross-sectional collaboration patterns, here we analyze the collaboration network from a researcher's local perspective along his/her career. Our longitudinal approach reveals that scientific collaboration is characterized by a high turnover rate juxtaposed with surprisingly frequent "life partners." We show that these extremely strong collaborations have a significant positive impact on productivity and citations—the apostle effect—representing the advantage of "super" social ties characterized by trust, conviction, and commitment.

Author contributions: A.M.P. designed research, performed research, analyzed data, and wrote the paper.

The author declares no conflict of interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence should be addressed. Email: petersen.xander@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1501444112/-/DCSupplemental.

results indicate that the "invisible college" defined by collaborative research activities (i.e., excluding informal communication channels and arm's-length associations) is surprisingly dominated by high-frequency interactions lasting only a few years. We then focus our analysis on the collaborative tie strength, K_{ij}, defined as the cumulative number of publications coauthored by i and j during the L_{ij} years of activity.

From the entire set of collaborators, we then identify a subset of super tie coauthors—those j with K_{ij} values that are statistically unlikely according to an author-specific extreme value criteria. Because almost all of the researchers we analyzed have more than one super tie, and roughly half of the publications we analyzed include at least one super tie coauthor, we were able to quantify the added value of super ties—for both productivity and citation impact—in two ways: (*i*) using descriptive measures and (*ii*) implementing a fixed-effects regression model. Controlling for author-specific features, we find that super ties are associated with increased publication rates and increased citation rates.

We term this finding the "apostle effect," signifying the dividends generated by extremely strong social ties based upon mutual trust, conviction, and commitment. This term borrows from biblical context, where an apostle represents a distinguished partner selected according to his/her noteworthy attributes from among a large pool of candidates. What we do not connote is any particular power relation (hierarchy) between i and the super tie coauthors, which is beyond the scope of this study. Also, because the perspective is centered around i, our super tie definition is not symmetric, i.e., if j is a super tie of i, i is not necessarily a super tie of j.

Because super ties have significant long-term impact on productivity and citations, our results are important from a career development perspective, reflecting the strategic benefits of cost, risk, and reward sharing via long-term partnership. The implications of research partnerships will become increasingly relevant as more careers become inextricably embedded in team science environments, wherein it can be difficult to identify contributions, signal achievement, and distribute credit. The credit distribution problem has received recent attention from the perspectives of institutional policy (8), team ethics (7), and practical implementation (27–29).

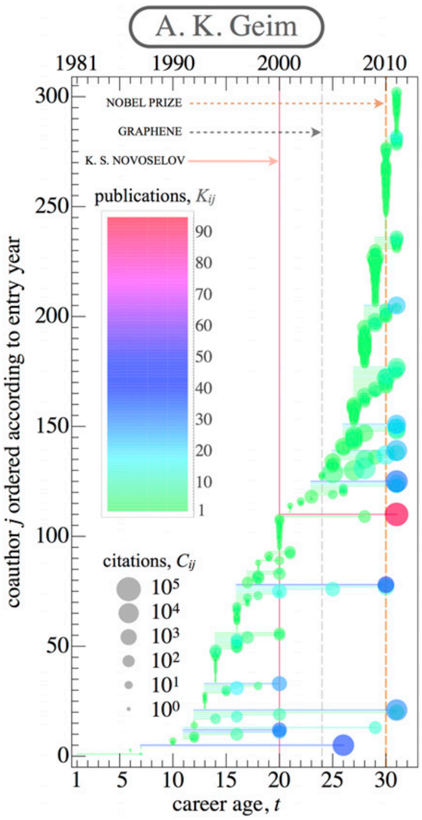

Fig. 1. Visualizing the embedding of academic careers in dynamic social networks. A career schematic showing A. Geim's collaborations, ordered by entry year. Notable career events include the first publication in 2000 with K. S. Novoselov (cowinner of the 2010 Nobel Prize in Physics) and their first graphene publication in 2004. An interesting network reorganization accompanies Geim's institutional move from Radboud University Nijmegen (The Netherlands) to University of Manchester (United Kingdom) in 2001. Moreover, the rapid accumulation of coauthors following the 2004 graphene discovery signals the new opportunities that accompany reputation growth.

Methods

Our study implements an ego network perspective, centered around each researcher career i, with weighted links connecting the central scientist to the peripheral nodes representing his/her collaborators (indexed by j). We constructed each ego network using longitudinal publication data from Thompson Reuters Web of Knowledge (TRWOK), comprising 193 biology and 280 physics careers in total. Each career profile is constructed by aggregating the publication, citation, and collaboration metadata over the first $t = 1 \ldots T_i$ years of his/her career. We downloaded the TRWOK data in calendar year Y_i, which is the citation count census year. Each disciplinary set includes a subset of 100 highly cited scientists (hereafter referred to as "Top"), selected using a ranking of the top-cited researchers in the high-impact journals *Physical Review Letters* and *Cell*. The rest of the researcher profiles (Other) are aggregated across physics and cell biology, with subsets that are specifically active in the domains of graphene, neuroscience, molecular biology, and genomics. The Other dataset only includes i with at least as many publications as the smallest N_i among the top-cited researchers: As such, $N_i \geq 52$ for biology and $N_i \geq 46$ for physics. This facilitates a reasonable comparison between Top and Other, possibly identifying differences attributable to innate success factors. See *SI Text* for further details on the data methods and selection.

This longitudinal approach leverages author-specific factors, revealing how career paths are affected by idiosyncratic events. To motivate this point, Fig. 1 illustrates the career trajectory of A. Geim, cowinner of the 2010 Nobel Prize in Physics. This schematic highlights three fundamental dimensions of collaboration ties—duration, strength, and impact: (*i*) each horizontal line indicates the collaboration of length $L_{ij} \equiv t^f_{ij} - t^0_{ij} + 1$ between i and coauthor j, beginning with their first joint publication in year t^0_{ij} and ending with their last observed joint publication in year t^f_{ij}; (*ii*) the circle color indicates the total number of joint publications, K_{ij}, representing our quantitative measure of tie strength; and (*iii*) the circle size indicates the net citations $C_{ij} = \sum_p c_{j,p}$ in Y_i, summed over the citations $c_{j,p}$ all publications p that include i and j.

This method of representing a science career, as illustrated in Figs. S1–S3, highlights the variability in collaboration strengths, both between and within career profiles. It is also worth mentioning that because multiple j may contribute to the same p, it is possible for coauthor measures to covary. However, for the remainder of the analysis, we focus on the dyadic relations between only i and j, leaving the triadic and higher-order team structures as an avenue for future work. For example, it would be interesting to know the likelihood of triadic closure between any two super ties of i, signaling coordinated cooperation; or, contrariwise, low triadic closure rates may indicate hierarchical organization around i.

Results

Quantifying the Collaboration Lifetime Distribution. We use L_{ij} to measure the duration of the productive interaction between i and j. Across researcher profiles, we find that a remarkable 60–80% of the collaborations have $L_{ij} = 1$ year (see Fig. S4). Considering the overwhelming dominance of the $L_{ij} = 1$ events, in this subsection, we concentrate our analysis on the subset of repeat collaborations with $L_{ij} > 1$ that produced two or more publications. Furthermore, due to censoring bias, L_{ij} values estimated for j who are active around the final career year of the data (T_i) may be biased toward small values. To account for this bias, in this subsection, we also exclude those collaborations that were active within the final L^c_i-year period, defining L^c_i as an initial average L_{ij} value calculated across all j for each i. Then, we calculate a second representative mean value, $\langle L_i \rangle$, which is calculated excluding the j with

$L_{ij} = 1$ and the j active in the final L_i^c-year period. Fig. 2*A* shows the probability distribution $P(\langle L_i \rangle)$, with mean values ranging from 4 y to 6 y, consistent with the typical duration of an early career position (e.g., PhD or postdoctoral fellow, assistant professor).

Establishing statistical regularities across research profiles requires the use of a normalized duration measure, $\Delta_{ij} \equiv L_{ij}/\langle L_i \rangle$, which controls for author-specific collaboration patterns by measuring time in units of $\langle L_i \rangle$. The empirical distributions are right-skewed, with approximately 63% of the data with $L_{ij} < \langle L_i \rangle$ (corresponding to $\Delta_{ij} < 1$). Nevertheless, ~1% of collaborations last longer than $4\langle L_i \rangle \approx 15$–$20$ y. Moreover, Fig. 2*A* shows that the log-logistic probability density function (pdf),

$$P(\Delta) = \frac{(b/a)(\Delta/a)^{b-1}}{\left(1+(\Delta/a)^b\right)^2}, \qquad [1]$$

provides a good fit to the empirical data over the entire range of Δ_{ij}. The log-logistic (Fisk) pdf is a well-known survival analysis distribution with property $\text{Median}(\Delta) = a$. By construction, the mean value $\langle \Delta \rangle \equiv 1$, which reduces our parameter space to just b as $a = \sin(\pi/b)/(\pi/b)$. For each dataset, we calculate $b \geq 2.6$, estimating the parameter using ordinary least squares. Associated with each $P(\Delta)$ is a hazard function representing the likelihood that a collaboration terminates for a given Δ_{ij}. Because $b > 1$, the hazard function is unimodal, with a maximum value occurring at $\Delta_c = a(b-1)^{1/b}$ with bounds $\Delta_c > a$ for $b > 2$ and $\Delta_c > 1$ for $b > 2.83...$; using the best-fit a and b values, we estimate $\Delta_c \approx 0.94$ (Top biology), 1.11 (Other biology), 0.77 (Top physics), and 1.08 (Other physics). Thus, Δ_c represents a tipping point in the sustainability of a collaboration, because the likelihood that a collaboration terminates peaks at Δ_c and then decreases monotonically for $\Delta_{ij} > \Delta_c$. This observation lends further significance to the author-specific time scale $\langle L_i \rangle$. The log-logistic pdf is also characterized by asymptotic power-law behavior $P(\Delta) \approx \Delta^{-(b+1)}$ for large Δ_{ij}.

To determine how the Δ_{ij} values are distributed across the career, we calculated the mean duration $\langle \Delta | t \rangle$ using a 5-y (sliding window) moving average centered around career age t. If the Δ_{ij} values were distributed independent of t, then $\langle \Delta | t \rangle \approx 1$. Instead, Fig. 2*B* shows a negative trend for each dataset. Interestingly, the $\langle \Delta | t \rangle$ values are consistently larger for the Top scientists, indicating that the relatively short L_{ij} are more concentrated at larger t. This pattern of increasing access to short-term collaboration opportunities points to an additional positive feedback mechanism contributing to cumulative advantage (30, 31).

Quantifying the Collaboration Life Cycle. The $P(\Delta)$ distribution points to the variability of time scales in the scientific collaboration network—although a small number of collaborations last a lifetime, the remainder decay quite quickly in a collaboration environment characterized by a remarkably high churn rate. Because it is possible that a relatively long L_{ij} corresponds to just the minimum two publications, it is also important to analyze the collaboration rate. To this end, we quantify the patterns of growth and decay in tie strength using the more than 166,000 dyadic (ij) collaboration records: $K_{ij}(t)$ is the cumulative number of coauthored publications between i and j up to year t, and $\Delta K_{ij}(t) = K_{ij}(t) - K_{ij}(t-1)$ is the annual publication rate.

To define a collaboration trajectory that is better suited for averaging, we normalize each individual $\Delta K_{ij}(\tau)$ by its peak value,

$$\Delta K'_{ij}(\tau) \equiv \Delta K_{ij}(\tau)/\text{Max}\left[\Delta K_{ij}(\tau)\right]. \qquad [2]$$

Here $\tau \equiv \tau_{ij} = t - t^0_{ij} + 1$ is the number of years since the initiation of a given collaboration. This normalization procedure is useful for comparing and averaging time series that are characterized by just a single peak.

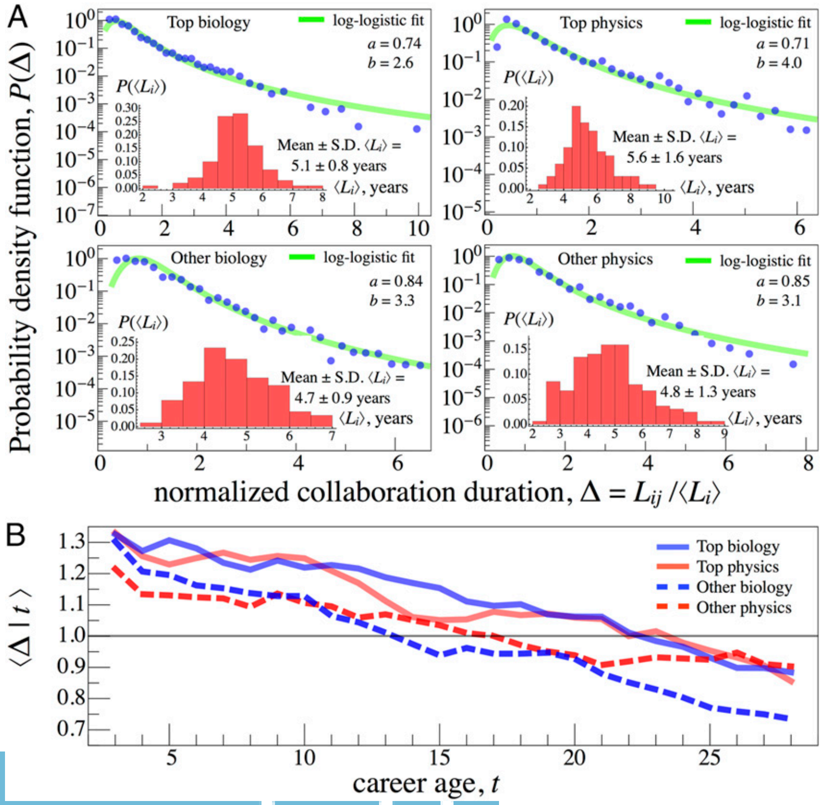

Fig. 2. Log-logistic distribution of collaboration duration. (*A*) The probability distribution $P(\Delta)$ is right-skewed and well fit by the log-logistic pdf defined in Eq. **1**. (*Insets*) The probability distribution $P(\langle L_i \rangle)$ shows that the characteristic collaboration length in physics and biology is typically between 2 y and 6 y. (*B*) The decrease in the typical collaboration timescale, $\langle \Delta | t \rangle$, reflects how careers transition from being pursuers of collaboration opportunities to attractors of collaboration opportunities.

Expecting that the collaboration trajectories depend on the tie strength, we grouped the individual $\Delta K'_{ij}(\tau)$ according to the normalized coauthor strength, $x_{ij} \equiv K_{ij}/\langle K_i \rangle$. The normalization factor $\langle K_i \rangle = S_i^{-1}\sum_{j=1}^{S_i} K_{ij}$ is calculated across the S_i distinct collaborators (the collaboration radius of i), and represents an intrinsic collaboration scale that grows in proportion to both an author's typical collaboration size and his/her publication rate. We then aggregated the $N_{\{x\}}$ trajectories in each $\{x\}$ group and calculated the average trajectory,

$$\left\langle \Delta K'_{ij}(\tau|x) \right\rangle \equiv N_{\{x\}}^{-1} \sum_{\{x\}} \Delta K'_{ij}(\tau|x). \qquad [3]$$

Indeed, Fig. 3 shows that the collaboration life cycle $\Delta K_{ij}(\tau|x)$ depends strongly on the relative tie strength $x_{ij} \equiv K_{ij}/\langle K_i \rangle$. The trajectories with $x_{ij} > 12.0$ decay over a relatively long time scale, maintaining a value approximately $0.2\text{Max}[\Delta K_{ij}(\tau)]$ even 20 y after initiation, reminiscent of a "research life partner." The trajectories with $x_{ij} \in [0.9, 1.4]$ represent common collaborations that decay exponentially over the characteristic time scale $\langle L_i \rangle$. A mathematical side note, useful as a modeling benchmark, is the linear decay when plotted on log-linear axes, suggesting a functional form that is exponential for large τ, $\langle \Delta K'_{ij}(\tau|x) \rangle \approx \exp[-\tau/\overline{\tau}]$.

We further emphasize the ramifications of the life cycle variation by quantifying the relation between x_{ij} and the collaboration's half-life $\tau_{1/2}$, defined as the number of years to reach half of the total collaborative output according to the relation $K_{ij}(t = \tau_{1/2}) = K_{ij}/2$. We observe a scaling relation for the average half life, $\langle \tau_{1/2} \rangle \approx x^{\zeta}$ with ζ values ranging from 0.4 to 0.5. Sublinear values ($\zeta < 1$) indicate that a collaboration with twice the strength is likely to have a corresponding $\tau_{1/2}$ that is less than doubled. This feature captures the burstiness of collaborative activities, which likely arises from the heterogenous overlapping of multiple timescales, e.g., the variable contract lengths in science ranging from single-year contracts to lifetime tenure, the overlapping of multiple age cohorts, and the projects and grants themselves, which are typically characterized by relatively short terms. Nevertheless, $dx/d\tau_{1/2} \approx \tau_{1/2}^{(1-\zeta)/\zeta}$ is increasing function for $\zeta < 1$, indicating an increasing marginal returns with increasing $\tau_{1/2}$, further signaling the productivity benefits of long-term collaborations characterized by formalized roles, mutual trust, experience, and group learning which together can facilitate efficient interactions.

Quantifying the Tie Strength Distribution. Here we focus on the cross-sectional distribution of tie strengths within the ego network. We use the final tie strength value K_{ij} to distinguish the strong ties ($K_{ij} \geq \langle K_i \rangle$) from the weak ties ($K_{ij} < \langle K_i \rangle$). Fig. 4*A* shows the cumulative distribution $P(\leq \langle K_i \rangle)$ of the mean tie strength $\langle K_i \rangle$, which can vary over a wide range depending on a researcher's involvement in large-team science activities. We also quantify the concentration of tie strength using the Gini index G_i calculated from each researcher's K_{ij} values; the distribution $P(\leq G_i)$ is shown in Fig. 4*B*. Together, these two measures capture the variability in collaboration strengths across and within disciplines, with physics exhibiting larger $\langle K_i \rangle$ and G_i values.

Another important author-specific variable is the publication overlap between each researcher and his/her top collaborator. This measure is defined as the fraction of a researcher's N_i publications including his/her top collaborator, $f_{K,i} = Max_j[K_{ij}]/N_i$. We observe surprisingly large variation in $f_{K,i}$, with mean and SD in the range of 0.16 ± 0.14 for the Top scientists and 0.36 ± 0.23 for the Other scientists. Across all profiles, the min and max $f_{K,i}$ values are 0.03 and 0.99, respectively, representing nearly the maximum possible variation in observed publication overlap. An example of this limiting scenario is shown in Fig. S2, highlighting the "dynamic duo" of J. L. Goldstein and M. S. Brown, winners of the 1985 Nobel Prize in Physiology or Medicine; Goldstein and Brown published more than 450 publications each, with roughly $100 \times f_{K,i} \approx 95\%$ coauthored together. Remarkably, we find that overlaps larger than 50% are not uncommon, observing $100P(f_K \geq 0.5) \approx 9\%$ (biology) and $100P(f_K \geq 0.5) \approx 20\%$ (physics) of i having more than half of their publications with their strongest collaborator.

However, within a researcher profile, it is likely that more than just the top collaborator was central to his/her career. Indeed, key to our investigation is the identification of the extremely strong collaborators—super ties—that are distinguished within the subset of strong ties. Hence, using the empirical information contained within each researcher's tie strength distribution, $P(K_{ij})$, we develop an objective super tie criteria that is author specific. First, to gain a better understanding of the statistical distribution of K_{ij}, we aggregated the tie strength data across all research profiles, using the normalized collaboration strength x_{ij}. Fig. 4 *C* and *D* shows the cumulative distribution $P(\geq x)$ for each discipline. Each $P(\geq x)$ is in good agreement with the exponential distribution $\exp[-x]$ (with mean value $\langle x \rangle = 1$ by construction), with the exception in the tail, $P(\geq x) \lesssim 10^{-3}$, which is home to extreme collaborator outliers. Thus, by a second means in addition to the result for L_{ij}, we find that roughly 2/3 of the ties we

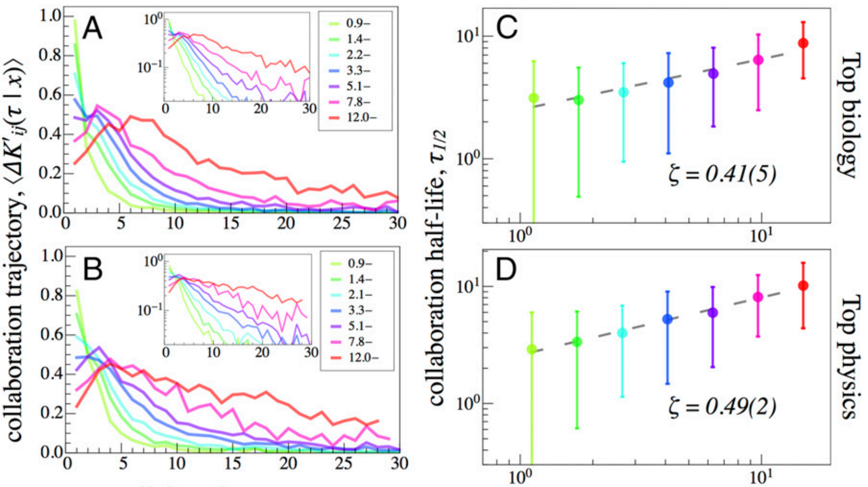

Fig. 3. Growth and decay of collaboration ties for (*A* and *C*) Top biology and (*B* and *D*) Top physics. (*A* and *B*) Average collaboration intensity, normalized to peak value, measured τ_{ij} years after the initiation of the collaboration tie. (*Insets*) On log-linear axes, the decay appears as 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 $x_{ij} \equiv K_{ij}/\langle K_i \rangle$. 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 relatively more productive, being characterized by increasing marginal returns ($1/\zeta > 1$). Fig. S3 shows the analogous plot for the Other physics and biology datasets; all four datasets exhibit similar features.

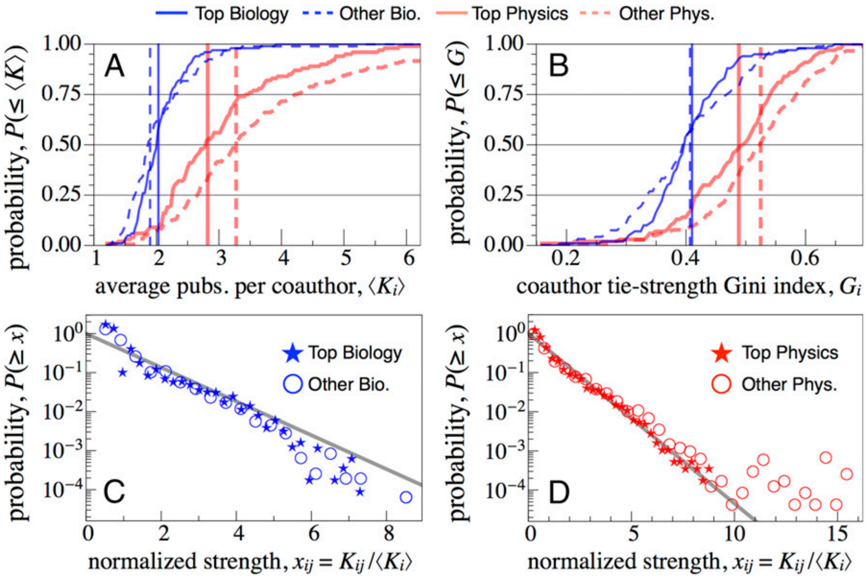

Fig. 4. Characteristic measures of collaboration tie strength. (*A*) Cumulative distribution of the mean collaboration strength, $\langle K_i \rangle$. The K-S test indicates that the $P(\langle K_i \rangle)$ are similar for biology ($p = 0.031$) and significantly different for physics ($p = 0.004$). Vertical lines indicate median value. (*B*) Cumulative distribution of G_i. The pairwise K-S test indicates that the $P(G_i)$ are similar for biology ($p = 0.14$) but not for physics ($p = 0.02$). Vertical lines indicate the mean value, with physics indicating significantly higher G_i than for biology. In (*C*) biology and (*D*) physics, for each dataset, the cumulative distribution of normalized collaboration strength x_{ij} shows excellent agreement with the exponential distribution $E(x) = \exp[-x]$ (gray line) over the bulk of the distribution, with the deviations in the tail regime representing less than 0.1% of the data.

analyzed are weak (i.e., the fraction of observations with $x_{ij} < 1$ is given by $1 - 1/e \approx 0.63$).

Based upon this empirical evidence, we use the discrete exponential distribution as our baseline model, $P(K_{ij}) \propto \exp(-\kappa_i K_{ij})$. We then use extreme statistics arguments to precisely define the author-specific super tie threshold K_i^c. The extreme statistic criterion posits that, out of the S_i empirical observations, there should be just a single observation with $K_{ij} > K_i^c$. The threshold K_i^c is operationalized by integrating the tail of $P(K_{ij})$ according to the equation $1/S_i = \sum_{K_{ij} > K_i^c}^{\infty} P(K_{ij}) = \exp(-\kappa_i K_i^c)$, with the analytic relation $\langle K_i \rangle = \sum_{K_{ij}=1}^{\infty} K_{ij} P(K_{ij}) = e^{\kappa_i}/(e^{\kappa_i} - 1) \approx 1 + 1/\kappa_i$ for small κ_i. In the relatively large S_i limit, K_i^c is given by the simple relation

$$K_i^c = (\langle K_i \rangle - 1) \ln S_i. \qquad [4]$$

The advantage of this approach is that K_i^c is nonparametric, depending only on the observables $\langle K_i \rangle$ and S_i. Thus, the super tie threshold is proportional to $\langle K_i \rangle - 1$ (the -1 arises because the minimum K_{ij} value is 1), with a logarithmic factor $\ln S_i$ reflecting the sample size dependence. This extreme value criteria is generic, and can be derived for any data following a baseline distribution; for a succinct explanation of this analytic method, see page 17 of ref. 32.

In what follows, we label each coauthor j with $K_{ij} > K_i^c$ a super tie, with indicator variable $R_j \equiv 1$. The rest of the ties with $K_{ij} \le K_i^c$ have an indicator variable $R_j \equiv 0$. This method has limitations, specifically in the case that the collaboration profile does not follow an exponential $P(K_{ij})$. For example, consider the extreme case where every $K_{ij} = 1$, meaning that $K_i^c = 0$ (independent of S_i), resulting in all coauthors being super ties ($R_j = 1$ for all j). This scenario is rare and unlikely to occur for researchers with relatively large N_i and S_i, as in our researcher sample.

Quantifying the Prevalence and Impact of Super Ties. How common are super ties? For each profile, we denote the number of coauthors that are super ties by $S_{R,i}$ (with complement $S_{!R,i} = S_i - S_{R,i}$). Fig. S4 shows that the distribution of $S_{R,i}$ is rather broad, with mean and SD $S_{R,i}$ values 18 ± 13 (Top biology), 16 ± 13 (Other biology), 7.3 ± 4.8 (Top physics), and 6.8 ± 5.1 (Other physics). The super tie coauthor fraction, $f_{R,i} = S_{R,i}/S_i$, measures the super tie frequency on a per-collaborator basis, with mean value $\langle f_R \rangle \approx 0.04$ (i.e., typically one super tie for every 25 coauthors). Furthermore, Fig. 5*A* shows that the distribution $P(\le f_R)$ is common across the four datasets. We tested the universality of the probability distribution $P(f_R)$ between the Top and Other researcher datasets using the Kolmogorov–Smirnov (K-S) statistic, which tests the null hypothesis that the data come from the same underlying pdf. The smallest pairwise K-S test *P* value between any two $P(f_R)$ is $p = 0.21$, indicating that we fail to reject the null hypothesis that the distributions are equal, highlighting that the four datasets are remarkably well matched with respect to the distribution of $f_{R,i}$.

On a per-paper basis, Fig. 5*B* shows that the fraction of a researcher's portfolio coauthored with at least one super tie, $f_{N,i}$, can vary over the entire range of possibilities, with mean and SD 0.50 ± 0.18 (Top biology), 0.74 ± 0.13 (Other biology), 0.42 ± 0.19 (Top physics), and 0.58 ± 0.23 (Other physics). Furthermore, we found that 41% of the Top scientists have $f_{N,i} \ge 0.5$. Interestingly, the distributions of $f_{K,i}$ and $f_{N,i}$ indicate that top scientists have lower levels of super tie dependency than their counterparts.

We also analyzed the arrival rate of super ties. For each profile, we tracked the number of super ties initiated in year t and normalized this number by the total number of new collaborations initiated in the same year. This ratio, $\lambda_{R,i}(t)$, estimates the likelihood that a new collaboration eventually becomes a super tie as a function of career age t. For example, using the set of collaborations initiated in each scientist's first year, we estimate the likelihood that a first-year collaborator (mentor) becomes a super tie at $\lambda_R(t = 1) = 8\%$ (Top biology), 16% (Other biology), 14% (Top physics), and 15% (Other physics). Fig. 5*D* shows the mean arrival rate, $\langle \lambda_R(t) \rangle$, calculated by averaging over all profiles in each dataset. The super tie arrival rate declines across the career, reaching a 5% likelihood per new collaborator at $t = 20$ and 2.5% likelihood by $t = 30$. The decay is not as fast for the top-cited scientists, possibly reflecting their preferential access to outstanding collaborators. However, the estimate for large t is biased toward smaller values because collaborations initiated late in the career may not have had sufficient time to grow.

In *The Apostle Effect I* and *The Apostle Effect II*, we investigate the role of super ties at the microlevel by analyzing productivity at the annual time resolution and the citation impact of individual publications. In *SI Text*, we provide additional evidence for the advantage of super ties by developing descriptive methods that measures the net productivity and citations of the super ties relative to all other ties.

The Apostle Effect I: Quantifying the Impact of Super Ties on Annual Productivity. We analyzed each research profile over the career

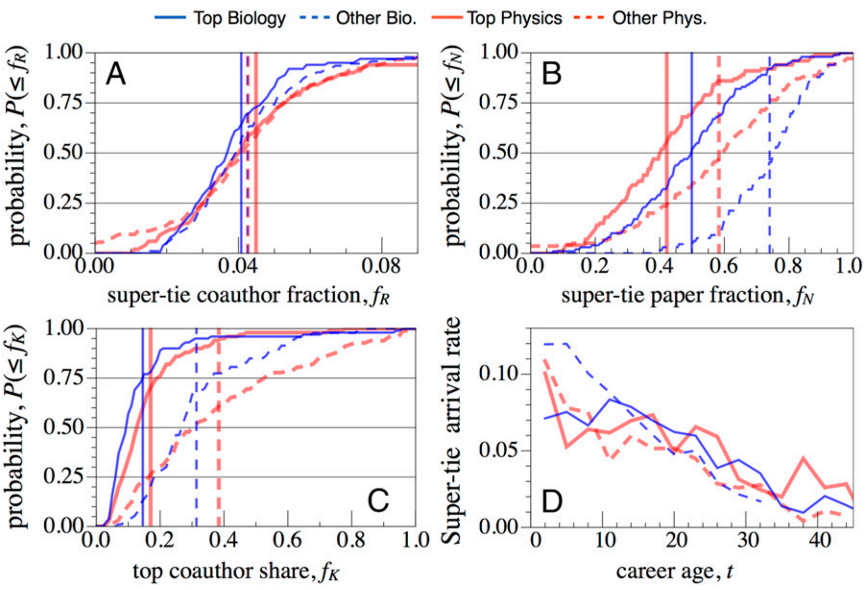

Fig. 5. The frequency of super ties. Vertical lines indicate the distribution mean. (*A*) Cumulative distribution of the fraction $f_{R,i}$ of the S_i coauthors that are super ties. All pairwise comparisons of the distributions have K-S *P* values greater than 0.21, indicating a common underlying distribution $P(f_R)$. (*B*) Cumulative distribution of the fraction $f_{N,i}$ of publications that include at least one super tie coauthor. The Top scientist distributions show mean values that are significantly smaller than their counterparts. (*C*) Cumulative distribution of the fraction $f_{K,i}$ of publications coauthored with his/her top collaborator. The mean and SD for biology (Top) is 0.15 ± 0.16, for biology (Other) is 0.31 ± 0.16, for physics (Top) is 0.17 ± 0.13, and for physics (Other) is 0.38 ± 0.26. (*D*) The mean rate of super ties per new collaboration, $\langle \lambda_R(t) \rangle$, averaged over all of the profiles in each dataset using observations aggregated over consecutive 3-y periods.

years $t_i \in [6, Min(29, T_i)]$, separating the data into nonoverlapping Δt-year periods, and neglecting the first 5 y to allow the $L_{ij}(t)$ and $K_{ij}(t)$ sufficient time to grow. We then modeled the dependent variable, $n_{i,t}/\langle n_i \rangle$, which is the productivity aggregated over Δt-year periods, normalized by the baseline average calculated over the period of analysis. Recent analysis of assistant and tenured professors has shown that the annual publication rate is governed by slow but substantial growth across the career, with fluctuations that are largely related to collaboration size (24).

To better understand the factors contributing to productivity growth, we include controls for career age t along with four additional variables measuring the composition of collaborators from each Δt-year period. First, we calculated the average number of authors per publication, $\overline{a}_{i,t}$, a proxy for labor input, coordination costs, and the research technology level. Second, we calculated the mean duration, $\overline{L}_{i,t}$, by averaging the $L_{ij}(t - \Delta t)$ values (from the previous period) across only the j who are active in t, i.e., those coauthors with $\Delta K_{ij}(t) > 0$. In this way, we account for the possibility that j was not active in the previous period $(t - \Delta t)$, in which case $L_{ij}(t - \Delta t)$ is even smaller than $L_{ij}(t) - \Delta t$. Thus, $\overline{L}_{i,t}$ measures the prior experience between i and his/her collaborators. Third, for the same set of coauthors as for $\overline{L}_{i,t}$, we calculated the Gini index of the collaboration strength, $G^K_{i,t}$, using the tie strength values up to the previous period, $K_{ij}(t - \Delta t)$. Thus, $G^K_{i,t}$ provides a standardized measure of the dispersion in coauthor activity, with values ranging from 0 (all coauthors published equally in the past with i) to 1 (extreme inequality in prior publication with i). Thus, whereas $\overline{L}_{i,t}$ measures the lifetime of the group's prior collaborations, $G^K_{i,t}$ measures the concentration of their prior experience. Finally, for each period t, we calculated the contribution of super tie collaborators normalized by the contribution of all other collaborators,

$$\rho_{i,t} \equiv \frac{\sum_{j|R=1} \Delta K_{ij}(t)}{\sum_{j|R=0} \Delta K_{ij}(t)}, \qquad [5]$$

accounting for the possibility that the relative contribution of super ties may affect productivity. Although the total coauthor contribution $\sum_j \Delta K_{ij}(t)$ is highly correlated with $n_{i,t}$, the correlation coefficient between $\rho_{i,t}$ and $n_{i,t}$ is only 0.07. We only include researchers in this analysis if there are ≥4 data points for which the denominator of Eq. **5** is nonzero.

We implemented a fixed-effects regression of the model

$$\frac{n_{i,t}}{\langle n_i \rangle} = \beta_{i,0} + \beta_{\overline{a}} \ln \overline{a}_{i,t} + \beta_{\overline{L}} \overline{L}_{i,t} + \beta_G G^K_{i,t} + \beta_\rho \rho_{i,t} + \beta_t t_{i,t} + \epsilon_{i,t}, \qquad [6]$$

which accounts for author-specific time-invariant features ($\beta_{i,0}$), using robust SEs to account for autocorrelation within each i. Because the predictors are calculated from the same ego profile, covariance is expected; for example, the highest correlation coefficient between any two independent variables is 0.32 between $\ln \overline{a}_{i,t}$ and $G^K_{i,t}$, because the variance in K_{ij} increases proportional to the sample size (i.e., $\overline{a}_{i,t}$). Table 1 shows the results of our model estimates for $\Delta t = 1$ year, and Table S1 shows the results for $\Delta t = 3$ years. We also ran the regression for all of the datasets together, "All," and provide standardized coefficients that better facilitate a comparison of the coefficient magnitudes.

We observed a positive coefficient $\beta_\rho = 0.11 \pm 0.01$ ($p \leq 0.003$ for all datasets), meaning that larger contributions by super ties are associated with above-average productivity. By way of example, consider a scenario where the super ties contribute a third of the total coauthor input, corresponding to $\rho_{i,t} = 0.5$, the average $\rho_{i,t}$ value we observed. Consider a second scenario with $\rho_{i,t} = 1$, corresponding to equal input by the super ties and their counterparts ($\rho_{i,t} \geq 1$ for 14% of the observations). If all other parameters contribute a baseline productivity value 1, then the additional contribution from β_ρ corresponds to a $100 \times 0.5\beta_\rho / (1 + 0.5\beta_\rho) = 5.2\%$ productivity increase. This value is consistent with the 5% productivity spillover observed in a study of star scientists (33).

We also found that periods corresponding to higher levels of prior experience are associated with below-average productivity ($\beta_{\overline{L}} < 0$, $p \leq 0.008$ for all datasets except for Top biology). Despite the costs associated with tie formation, this result demonstrates that productivity can benefit from collaborator turnover. Nevertheless, above-average productivity is associated with higher inequality in

Table 1. Parameter estimates for the productivity model for $n_{i,t}$ in Eq. 6 using $\Delta t = 1$-y-long periods

Dataset	A	$\ln \bar{a}_t$	$\bar{L}_t$	G_t^K	ρ_t	t	$N_{obs.}$	Adj. R^2
All	466	0.002 ± 0.029	**−0.054 ± 0.008**	**1.788 ± 0.134**	**0.110 ± 0.013**	**0.029 ± 0.002**	8,483	0.19
(Std. coeff.)		0.002 ± 0.033	**−0.140 ± 0.021**	**0.320 ± 0.024**	**0.140 ± 0.016**	**0.049 ± 0.004**		
P value		0.943	0.000	0.000	0.000	0.000		
Biology (Top)	99	**−0.123±0.056**	−0.011 ± 0.018	**2.816 ± 0.270**	**0.111 ± 0.026**	**0.031 ± 0.003**	2,202	0.24
P value		0.031	0.519	0.000	0.000	0.000		
Biology (Other)	95	−0.061 ± 0.056	**−0.067 ± 0.025**	**1.654 ± 0.287**	**0.071 ± 0.023**	**0.053 ± 0.006**	1,467	0.29
P value		0.275	0.008	0.000	0.003	0.000		
Physics (Top)	100	**−0.146 ± 0.057**	**−0.047 ± 0.015**	**2.053 ± 0.287**	**0.153 ± 0.025**	**0.022 ± 0.004**	2,056	0.15
P value		0.012	0.002	0.000	0.000	0.000		
Physics (Other)	172	0.089 ± 0.050	**−0.065 ± 0.013**	**1.495 ± 0.213**	**0.101 ± 0.021**	**0.026 ± 0.005**	2,758	0.15
P value		0.079	0.000	0.000	0.000	0.000		

Each fixed-effects model was calculated using robust SEs, implemented by the Huber/White/sandwich method. Values significant at the $p \leq 0.04$ level are indicated in boldface. Std. coeff., the estimates of the standardized (beta) coefficients; All, the combination of all datasets.

the concentration of prior experience ($\beta_G > 0, p < 0.001$ level for all datasets). Together, these results point to the benefits of strategically pairing new collaborators with incumbent ones to promote the atypical combination of knowledge backgrounds and to achieve higher scientific impact (34). The standardized coefficients in Table 1 indicate that β_G is twice as strong as β_ρ and $\beta_{\bar{L}}$; interestingly, β_ρ and $\beta_{\bar{L}}$ have opposite signs yet are balanced in magnitude, suggesting a compensation strategy for group managers.

The age coefficient β_t is also positive ($p < 0.001$ level for all datasets), consistent with patterns of steady productivity growth observed for successful research careers (5, 24, 31). Possible explanatory variables to consider in extended analyses are the SD in K_{ij}, a contact frequency (K_{ij}/L_{ij}) measure of tie strength intensity per Granovetter's original operationalization (10), and absolute calendar year y, variables that we omit here to keep the model streamlined.

The Apostle Effect II: Quantifying the Impact of Super Ties on the Long-Term Citation of Individual Publications. The impact of super ties on a publication's long-term citation tally is difficult to measure, because, clearly, older publications have had more time to accrue citations than newer ones—a type of censoring bias—and so a direct comparison of raw citation counts for publications from different years is technically flawed. To address this measurement problem, we map each publication's citation count $c_{i,p,Y}(y)$ in census year Y_i to a normalized z score,

$$z_{i,p,y} \equiv \frac{\ln c_{i,p,Y}(y) - \langle \ln c_Y^m(y) \rangle}{\sigma[\ln c_Y^m(y)]}. \qquad [7]$$

This citation measure is well suited for the comparison of publications from different y because $z_{i,p,y}$ is measured relative to the mean $\langle \ln c_Y^m(y) \rangle$ number of citations by publications from the same year y, in units of the SD, $\sigma[\ln c_Y^m(y)]$ (31). Thus, we take advantage of the fact that the distribution of citations obeys a universal log-normal distribution for p from the same y and discipline (35). In this way, z is defined such that the distribution $P(z)$ is sufficiently time invariant. To confirm this property, we aggregated $z_{i,p,y}$ within successive 8-y periods, and calculated the conditional distributions $P(z|y)$, which are stable and approximately normally distributed over the entire sample period (Fig. S5).

To define the detrending indices $\langle \ldots \rangle$ and $\sigma[\ldots]$, we use the baseline journal set m comprising all research articles collected from the journals *Nature*, *Proceedings of the National Academy of Sciences*, and *Science*. We use this aggregation of three multidisciplinary journals only to control for the time-dependent feature of citation counts. We chose these journals as our baseline because they have relatively large impact factors (high citation rates), and so the temporal information contained in $\langle \ldots \rangle$ and $\sigma[\ldots]$ is less noisy than other m with lower citation rates. Furthermore, because most publications reach their peak citation rate within 5–10 y after publication (5), we only analyze $z_{i,p,y}$ with $y \leq 2003$. In this way, the $z_{i,p,y}$ values we analyze are less sensitive to fluctuations early in the citation lifecycle, in addition to recent paradigm shifts in science such as the Internet, which affects the search, the retrieval, and the citation of prior literature, and the rise of open access publishing.

In our regression model, we use five explanatory variables that are author (i) and publication (p) specific. The first is the number of coauthors, $a_{i,p}$, which controls for the tendency for publications with more coauthors to receive more citations (4). This variable is also a gross level of technology and coordination costs, because larger teams typically reflect endeavors with higher technical challenge distributed across a wider range of skill sets. We use $\ln a_{i,p}$ because the range of values is rather broad, appearing to be

Table 2. Parameter estimates for the citation model for $z_{i,p}$ in Eq. 8 using only the publications with $y_p \leq 2003$

Dataset	A	$\ln a_p$	R_p	t_p	$\ln N_i(t_p)$	$\ln S_i(t_p)$	$N_{obs.}$	Adj. R^2
All	377	**0.263 ± 0.024**	**0.202 ± 0.023**	**−0.061 ± 0.004**	0.062 ± 0.066	0.065 ± 0.072	68,589	0.27
(Std. coeff.)		**0.135 ± 0.012**	**0.129 ± 0.015**	**−0.039 ± 0.003**	0.044 ± 0.046	0.050 ± 0.055		
P value		0.000	0.000	0.000	0.347	0.367		
Biology (Top)	100	**0.263 ± 0.039**	**0.213 ± 0.033**	**−0.029 ± 0.007**	−0.138 ± 0.102	0.062 ± 0.112	22,135	0.12
P value		0.000	0.000	0.000	0.177	0.578		
Biology (Other)	55	**0.579 ± 0.053**	**0.152 ± 0.066**	**−0.031 ± 0.015**	−0.179 ± 0.095	**0.211 ± 0.094**	4,801	0.20
P value		0.000	0.026	0.040	0.065	0.029		
Physics (Top)	100	**0.139 ± 0.043**	**0.230 ± 0.044**	**−0.070 ± 0.007**	**0.277 ± 0.118**	−0.119 ± 0.135	22,673	0.19
P value		0.002	0.000	0.000	0.021	0.380		
Physics (Other)	122	**0.272 ± 0.042**	**0.235 ± 0.049**	**−0.060 ± 0.008**	0.082 ± 0.095	0.017 ± 0.104	18,980	0.19
P value		0.000	0.000	0.000	0.389	0.870		

Each fixed-effects model was calculated using robust SEs, implemented by the Huber/White/sandwich method. Values significant at the $p \leq 0.04$ level are indicated in boldface. Std. coeff., the estimates of the standardized (beta) coefficients; All, the combination of all datasets.

approximately log-normally distributed in the right tail (7). The second explanatory variable is the dummy variable $R_{i,p}$, which takes the value 1 if p includes a super tie and the value 0 otherwise. Remarkably, the percentage of publications including a super tie is rather close to parity for three of the four datasets: 54% (Top biology), 45% (Top physics), 74% (Other biology), and 54% (Other physics). The third age variable, $t_{i,p}$, is the career age of i at the time of publication. The fourth variable, $N_i(t_p)$, is the total number of publications up to year $t_{i,p}$, which is a non-citation-based measure of the central author's reputation, visibility, and experience within the scientific community. The final explanatory variable is the collaboration radius, $S_i(t_p)$, which is the cumulative number of distinct coauthors up to $t_{i,p}$, representing the central author's access to collaborative resources, as well as an estimate of the number of researchers in the local community who, having published with i, may preferentially cite i. Hence, by including $N_i(t_p)$ and $S_i(t_p)$, we control for two dimensions of cumulative advantage that could potentially affect a publication's citation tally.

We then implement a fixed-effects regression to estimate the parameters of the citation impact model,

$$z_{i,p}=\beta_{i,0}+\beta_a \ln a_{i,p}+\beta_R R_{i,p}+\beta_t t_{i,p}+\beta_N \ln N_i(t_p)+\beta_S \ln S_i(t_p)+\epsilon_{i,p}, \qquad \textbf{[8]}$$

using the Huber/White/sandwich method to calculate robust SE estimates that account for heteroskedasticity and within-panel serial correlation in the idiosyncratic error term $\epsilon_{i,p}$. We excluded publications with $y_p > 2003$, and, in order that the Top and Other datasets are well balanced, we also excluded the Other researchers with less than 43 (biology) and 33 (physics) publications (observations) as of 2003. Table 2 lists the (standardized) parameter estimates. We provide the data used for both regression models in Dataset S1.

We estimated $\beta_R = 0.20 \pm 0.02$ ($p \leq 0.026$ level in each regression), indicating a significant relative citation increase when a publication is coauthored with at least one super tie. The standardized β_a and β_R coefficients are roughly equal, meaning that increasing a_p from 1 (a solo author publication) to $e \approx 3$ coauthors produces roughly the same effect as a change in R_p from 0 to 1. Thus, although larger team size correlates with more citations (4), the relative strength of β_R stresses the importance of who in addition to how many.

Interestingly, the career age parameter $\beta_t = -0.061 \pm 0.004$ is negative (significant at the $p \leq 0.04$ level in each regression), meaning that researchers' normalized citation impact decreases across the career, possibly due to finite career and knowledge life cycles. This finding is consistent with a large-scale analysis of researcher histories within high-impact journals, which also shows a negative trend in the citation impact across a career (31). Neither the reputation (β_N) nor collaboration radius (β_S) parameters were consistently statistically significant in explaining $z_{i,p,y}$, likely because they are highly correlated with t_p for established researchers. Modifications to consider in followup analysis are controls for the impact factor of the journal publishing p, the absolute year y to account for shifts in citation patterns in the post-Internet era, and removing self-citations from super ties. Unfortunately, this last task requires a substantial increase in data coverage, far beyond the relatively small amount needed to construct individual ego network collaboration profiles.

We develop three additional descriptive methods in *SI Text* to compare the subset of publications with at least one super tie to the complementary subset of publications without one. These investigations provide further evidence for the apostle effect. First, we defined an aggregate career measure, the productivity premium $p_{N,i}$ (see Eq. **S1**), which measures the average K_{ij} value among the super ties relative to all of the other collaborators. Second, we defined a similar career measure, the citation premium $p_{C,i}$ (see Eq. **S5**), which quantifies the average citation impact attributable to super ties relative to all of the other collaborators.

Independent of dataset, we observed rather substantial premium values. For example, the productivity premium has an average value $\langle p_N \rangle \approx 8$, meaning that on a per-collaborator basis, productivity with super ties is roughly 8 times higher than with the remaining collaborators. Similarly, the citation premium $p_{C,i}$ is also significantly right-skewed, with average value $\langle p_C \rangle \approx 14$, meaning that net citation impact per super tie is 14 times larger than the net citation impact from all other collaborators. We emphasize that $p_{C,i}$ appropriately accounts for team size by using an equal partitioning of citation credit across the a_p coauthors, remedying the multiplicity problem concerning citation credit.

Third, we calculated an additional estimation of the publication-level citation advantage due to super ties (Fig. S6). For both biology and physics, we found that the publications with super ties receive roughly 17% more citations than their counterparts. In basic terms, this means that the average publication with a super tie has 21 more citations in biology and 8 more citations in physics than the average publication without a super tie. This is not a tail effect, because the citation boost factor $\alpha_R = 1.17$ applies a multiplicative shift to the entire citation distribution, $P(\tilde{c}|R_p = 1) \approx P(\alpha_R \tilde{c}|R_p = 0)$, thereby impacting publications above and below the average.

Discussion

The characteristic collaboration size in science has been steadily increasing over the last century (4, 7, 21), with consequences at every level of science, from education and academic careers to universities and funding bodies (8). Understanding how this team-oriented paradigm shift affects the sustainability of careers, the efficiency of the science system, and society's capacity to overcome grand challenges will be of great importance to a broad range of scientific actors, from scientists to science policy makers.

Collaborative activities are also fundamental to the career growth process, especially in disciplines where research activities require a division of labor. This is especially true in biology and physics research, where computational, theoretical, and experimental methods provide complementary approaches to a wide array of problems. As a result, a contemporary research group leader is likely to find the assembly of team—one that is composed of individuals with diverse yet complementary skill sets—a daunting task, especially when under constraints to optimize financial resources, valuable facilities, and other material resources. Online social network platforms, such as VIVO (www.vivoweb.org/) and Profiles RNS (profiles.catalyst.harvard.edu/), which serve as match-making recommendation systems, have been developed to facilitate the challenges of team assembly.

Our analysis indicates that 2/3 of the collaborations analyzed here are weak. Nevertheless, the remaining strong ties represent social capital investments that can indeed have important long-term implications, for example, on information spreading (17), career paths (36), and access to key strategic resources (37). In the private sector, strong ties facilitate access to new growth opportunities, playing an important role in sustaining the competitiveness of firms and employees (38). These considerations further identify why it is important for researchers to understand the opportunities that exist within their local network. Understanding the redundancies in the local network (39) and the interaction capacity of team members (25) can help a group leader optimize group intelligence (26) and monitor team efficiency (24), thereby constituting a source of strategic competitive advantage.

In summary, we developed methods to better understand the diversity of collaboration strengths. We focused on the career as the unit of analysis, operationalized by using an ego perspective so that collaborations, publications, and impact scores fit together into a temporal framework ideal for cross-sectional and longitudinal modeling. Analyzing more than 166,000 collaborations, we found that a remarkable 60–80% of the collaborations last only $L_{ij}=1$ year. Within a subset of repeat collaborations ($L_{ij}\geq$ 2 y), we find that roughly 2/3 of these collaborations last less than a scientist's average duration $\langle L_i\rangle \approx 5$ y, yet 1% last more than $4\langle L_i\rangle \approx 20$ y. This wide range in duration and the disparate frequencies of long and short L_{ij} together point to the dichotomy of burstiness and persistence in scientific collaboration. Closer inspection of individual career paths signals how idiosyncratic events, such as changing institutions or publishing a seminal study or book, can have significant downstream impact on the arrival rate of new collaboration opportunities and tie formation (see Fig. 1 and Fig. S1). Also, the frequency of relatively large publication overlap measures ($f_{K,i}$ and $f_{N,i}$) indicates that career partners occur rather frequently in science.

In the first part of the study, we provided descriptive insights into basic questions such as how long are typical collaborations, how often does a scientist pair up with his/her main collaborator, and what is the characteristic half-life of a collaboration. We also found that as the career progresses, researchers become attractors rather than pursuers of new collaborations. This attractive potential can contribute to cumulative advantage (30, 31), as it provides select researchers access to a large source of collaborators, which can boost productivity and increase the potential for a big discovery.

We operationalized tie strength using an egocentric perspective of the collaboration network. Because the number of publications K_{ij} between the central scientist i and a given coauthor j was found to be exponentially distributed, the mean value $\langle K_i\rangle$ is a natural author-specific threshold that distinguishes the strong ($K_{ij}\leq\langle K_i\rangle$) from the weak ties ($K_{ij}<\langle K_i\rangle$). Within the subset of strong ties, we identified super tie outliers using an analytic extreme-statistics threshold K_i^c defined in Eq. **4**. Also, because the number of publications produced by a collaboration is highly correlated with its duration, a super tie also represents persistence that is in excess of the stochastic churn rate that is characteristic of the scientific system. On a per-collaborator basis, the fraction of coauthors within a research profile that are super ties ($f_{R,i}$) was remarkably common across datasets, indicating that super ties occur at an average rate of 1 in 25 collaborators.

There are various candidate explanations for why such extremely strong collaborations exist. Prosocial motivators may play a strong role, i.e., for some researchers, doing science in close community may be more rewarding than going it alone. Also, the search and formation of a compatible partnership requires time and other social capital investment, i.e., networking. Hence, for two researchers who have found a collaboration that leverages their complementarity, the potential benefits of improving on their match are likely outweighed by the long-term returns associated with their stable partnership. Complementarity, and the greater skill set the partnership brings, can also provide a competitive advantage by way of research agility, whereby a larger collective resource base can facilitate rapid adjustments to new and changing knowledge fronts, thereby balancing the risks associated with changing research direction. After all, a first-mover advantage can make a significant difference in a winner-takes-all credit and reward system (2).

Scientists may also strategically pair up to share costs, rewards, and risk across their careers. In this light, an additional incentive to form super ties may be explained, in part, by the benefits of reward sharing in the current scientific credit system, wherein publication and citation credit arising from a single publication are multiplied across the a_p coauthors in everyday practice. Considered in this way, the career risk associated with productivity lulls can be reduced if a close partnership is formed. For example, we observed a few "twin profiles" characterized by a publication overlap fraction $f_{K,i}$ between the researcher and his/her top collaborator that was nearly 100%. Moreover, we found that 9% of the biologists and 20% of the physicists shared 50% or more of their papers with their top collaborator. This highlights a particularly difficult challenge for science, which is to develop a credit system that appropriately divides the net credit but, at the same time, does not reduce the incentives for scientists to collaborate (8, 27–29). Thus, it will be important to consider these relatively high levels of publication and citation overlap in the development of quantitative career evaluation measures; otherwise, there is no penalty to discourage coauthor free riding (7).

We concluded the analysis by implementing two fixed-effects regression models to determine the sign and strength of the apostle effect represented by β_ρ (productivity) and β_R (citations). Together, these two coefficients address the fundamental question: Is there a measurable advantage associated with heavily investing in a select group of research partners?

In the first model, we measured the impact of super ties on a researcher's annual publication rate, controlling for career age, average team size, the prior experience of i with his/her coauthors, and the relative contribution of super ties within year t as measured by $\rho_{i,t}$ in Eq. **5**. We found larger $\rho_{i,t}$ to be associated with above-average productivity ($\beta_\rho>0$), indicating that super ties play a crucial role in sustaining career growth. We also found increased levels of prior experience to be associated with decreased productivity ($\beta_{\overline{L}}<0$), suggesting that maintaining older ties conflicts with the potential benefits from mixing new collaborators into the environment. Nevertheless, higher inequality in the concentration of prior experience was found to have a counterbalancing positive effect on productivity ($\beta_G>0$).

In the second regression model, we analyzed the impact of super ties on the citation impact of individual publications, using the detrended citation measure $z_{i,p,y}$ defined in Eq. **7**. This citation measure is normalized within publication year cohorts, thus allowing for a comparison of citation counts for research articles published in different years. We found that publications coauthored with super ties, corresponding to 52% of the papers we analyzed, have a significant increase in their long-term citations ($\beta_R>0$). In *SI Text*, we provide additional evidence for the apostle effect, showing that publications with super ties receive 17% more citations. This added value may arise from the extra visibility the publications receives, because the super tie collaborator may also contribute a substantial reputation and future productivity that promote the visibility of the publication. This type of network-mediated reputation spillover is corroborated by a recent study finding a significant citation boost attributable to a researcher's centrality within the collaboration network (40).

This data-oriented analysis also contributes to the literature on the science of science policy (41), providing insight and guidance in an increasingly metrics-based evaluation system on how to account for individual achievement in team settings. As such, we conclude with some policy recommendations. One particularly relevant scenario is fellowship, tenure, and career award evaluations, where it is a common practice to consider "independence from one's thesis advisor" as a selection criteria. We show that to assess a researcher's independence, evaluation committees should also take into consideration the level of publication overlap between a researcher and his/her strongest collaborator(s), e.g., $f_{K,i}$ and $f_{N,i}$. However, at the same time, the beneficial role of super ties—as we have quantitatively demonstrated—should also be acknowledged and supported. For example, funding programs might consider career awards that are specifically multipolar (8), which would also benefit the research partners in academia who are actually life partners, and who may face the daunting "two-body problem" of coordinating two research careers. Furthermore, understanding the basic levels of publication overlap in science is also important for the ex post facto review of funding outcomes as a means to evaluate the efficiency of science. In large-team settings, measuring the efficiency of a laboratory or project is difficult without a better understanding of how to measure overlapping labor inputs (i.e., collaborator contributions) relative to the project outputs (e.g., publications, patents, etc.). Finally, our study informs early career researchers—who are likely to face important decisions concerning the (possibly strategic) selection of collaborative opportunities—on the positive impact that the right research partner can have on their career's long-term sustainability and growth. In all, our results provide quantitative insights into the benefits associated with strong collaborative partnerships, pointing to the added value derived from skill-set complementarity, social trust, and long-term commitment.

ACKNOWLEDGMENTS. The author is grateful for helpful discussions with O. Doria, M. Imbruno, B. Tuncay, and R. Metulini and constructive criticism and keen insights from two anonymous referees. The author also acknowledges feedback from participants in the European Union Cooperation in Science and Technology (COST) Action TD1210 (KnowEscape) workshop on "Quantifying scientific impact: Networks, measures, insights?" and support from the Italian Ministry of Education for the National Research Project (PNR) "Crisis Lab."

1. Börner K, et al. (2010) A multi-level systems perspective for the science of team science. *Sci Transl Med* 2(49):49cm24.
2. Stephan P (2012) *How Economics Shapes Science* (Harvard Univ Press, Cambridge, MA).
3. Nahapiet J, Ghoshal S (1998) Social capital, intellectual capital, and the organizational advantage. *Acad Manage Rev* 23(2):242–266.
4. Wuchty S, Jones BF, Uzzi B (2007) The increasing dominance of teams in production of knowledge. *Science* 316(5827):1036–1039.
5. Petersen AM, et al. (2014) Reputation and impact in academic careers. *Proc Natl Acad Sci USA* 111(43):15316–15321.
6. Malmgren RD, Ottino JM, Nunes Amaral LA (2010) The role of mentorship in protégé performance. *Nature* 465(7298):622–626.
7. Petersen AM, Pavlidis I, Semendeferi I (2014) A quantitative perspective on ethics in large team science. *Sci Eng Ethics* 20(4):923–945.
8. Pavlidis I, Petersen AM, Semendeferi I (2014) Together we stand. *Nat Phys* 10:700–702.
9. Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. *Science* 323(5916):892–895.
10. Granovetter MS (1973) The strength of weak ties. *Am J Sociol* 78(6):1360–1380.
11. Newman MEJ (2001) The structure of scientific collaboration networks. *Proc Natl Acad Sci USA* 98(2):404–409.
12. Newman MEJ (2001) Scientific collaboration networks. I. Network construction and fundamental results. *Phys Rev E Stat Nonlin Soft Matter Phys* 64(1 Pt 2):016131.
13. Barabasi AL, et al. (2002) Evolution of the social network of scientific collaborations. *Physica A* 311(34):590–614.
14. Newman MEJ (2004) Coauthorship networks and patterns of scientific collaboration. *Proc Natl Acad Sci USA* 101(Suppl 1):5200–5205.
15. Guimerà R, Uzzi B, Spiro J, Amaral LAN (2005) Team assembly mechanisms determine collaboration network structure and team performance. *Science* 308(5722):697–702.
16. Palla G, Barabási AL, Vicsek T (2007) Quantifying social group evolution. *Nature* 446(7136):664–667.
17. Pan RK, Saramäki J (2012) The strength of strong ties in scientific collaboration networks. *Europhys Lett* 97(1):18007.
18. Martin T, Ball B, Karrer B, Newman MEJ (2013) Coauthorship and citation patterns in the Physical Review. *Phys Rev E Stat Nonlin Soft Matter Phys* 88(1):012814.
19. Ke Q, Ahn YY (2014) Tie strength distribution in scientific collaboration networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 90(3):032804.
20. Börner K, Maru JT, Goldstone RL (2004) The simultaneous evolution of author and paper networks. *Proc Natl Acad Sci USA* 101(Suppl 1):5266–5273.
21. Milojević S (2014) Principles of scientific research team formation and evolution. *Proc Natl Acad Sci USA* 111(11):3984–3989.
22. March JG (1991) Exploration and exploitation in organizational learning. *Organ Sci* 2(1):71–87.
23. Lazer D, Friedman A (2007) The network structure of exploration and exploitation. *Adm Sci Q* 52(4):667–694.
24. Petersen AM, Riccaboni M, Stanley HE, Pammolli F (2012) Persistence and uncertainty in the academic career. *Proc Natl Acad Sci USA* 109(14):5213–5218.
25. Pentland A (2012) The new science of building great teams. *Harv Bus Rev* 90:60–69.
26. Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW (2010) Evidence for a collective intelligence factor in the performance of human groups. *Science* 330(6004):686–688.
27. Stallings J, et al. (2013) Determining scientific impact using a collaboration index. *Proc Natl Acad Sci USA* 110(24):9680–9685.
28. Allen L, Scott J, Brand A, Hlava M, Altman M (2014) Publishing: Credit where credit is due. *Nature* 508(7496):312–313.
29. Shen HW, Barabási AL (2014) Collective credit allocation in science. *Proc Natl Acad Sci USA* 111(34):12325–12330.
30. Petersen AM, Jung WS, Yang JS, Stanley HE (2011) Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proc Natl Acad Sci USA* 108(1):18–23.
31. Petersen AM, Penner O (2014) Inequality and cumulative advantage in science careers: A case study of high-impact journals. *EPJ Data Sci* 3:24.
32. Krapivsky P, Redner S, Ben-Naim E (2010) *A Kinetic View of Statistical Physics* (Cambridge Univ Press, Cambridge, UK).
33. Azoulay P, Zivin JSG, Wang J (2010) Superstar extinction. *Q J Econ* 125(2):549–589.
34. Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. *Science* 342(6157):468–472.
35. Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: Toward an objective measure of scientific impact. *Proc Natl Acad Sci USA* 105(45):17268–17272.
36. Clauset A, Arbesman S, Larremore DB (2015) Systematic inequality and hierarchy in faculty hiring networks. *Sci Adv* 1(1):e1400005.
37. Duch J, et al. (2012) The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PLoS One* 7(12):e51332.
38. Uzzi B (1999) Embeddedness in the making of financial capital: How social relations and networks benefit firms seeking financing. *Am Sociol Rev* 64(4):481–505.
39. Burt RS (1992) *Structural Holes* (Harvard Univ Press, Cambridge, MA).
40. Sarigl E, Pfitzner R, Scholtes I, Garas A, Schweitzer F (2014) Predicting scientific success based on coauthorship networks. *EPJ Data Sci* 3:9.
41. Fealing KH, ed (2011) *The Science of Science Policy: A Handbook* (Stanford Business Books, Stanford, CA).
42. Petersen AM, Jung WS, Stanley HE (2008) On the distribution of career longevity and the evolution of home run prowess in professional baseball. *Europhys Lett* 83(5):50010.
43. Petersen AM, Penner O, Stanley HE (2011) Methods for detrending success metrics to account for inflationary and deflationary factors. *Eur Phys J B* 79(1):67–78.
44. Petersen AM, Wang F, Stanley HE (2010) Methods for measuring the citations and productivity of scientists across time and discipline. *Phys Rev E Stat Nonlin Soft Matter Phys* 81(3 Pt 2):036114.
45. Petersen AM, Stanley HE, Succi S (2011) Statistical regularities in the rank-citation profile of scientists. *Sci Rep* 1:181.
46. Petersen AM, Succi S (2013) The Z-index: A geometric representation of productivity and impact which accounts for information in the entire rank-citation profile. *J Informetrics* 7(4):823–832.
47. Penner O, Pan RK, Petersen AM, Kaski K, Fortunato S (2013) On the predictability of future impact in science. *Sci Rep* 3:3052.
48. Acemoglu D, Robinson JA (2005) *Economic Origins of Dictatorship and Democracy* (Cambridge Univ Press, Cambridge, UK).
49. Ausloos M (2013) A scientometrics law about co-authors and their ranking: The co-author core. *Scientometrics* 95:895–909.