

Principles of scientific research team formation and evolution

Staša Milojević¹

School of Informatics and Computing, Indiana University, Bloomington, IN 47401

Edited by Yu Xie, University of Michigan, Ann Arbor, MI, and approved December 30, 2013 (received for review May 22, 2013)

Research teams are the fundamental social unit of science, and yet there is currently no model that describes their basic property: size. In most fields, teams have grown significantly in recent decades. We show that this is partly due to the change in the character of team size distribution. We explain these changes with a comprehensive yet straightforward model of how teams of different sizes emerge and grow. This model accurately reproduces the evolution of empirical team size distribution over the period of 50 y. The modeling reveals that there are two modes of knowledge production. The first and more fundamental mode employs relatively small, “core” teams. Core teams form by a Poisson process and produce a Poisson distribution of team sizes in which larger teams are exceedingly rare. The second mode employs “extended” teams, which started as core teams, but subsequently accumulated new members proportional to the past productivity of their members. Given time, this mode gives rise to a power-law tail of large teams (10–1,000 members), which features in many fields today. Based on this model, we construct an analytical functional form that allows the contribution of different modes of authorship to be determined directly from the data and is applicable to any field. The model also offers a solid foundation for studying other social aspects of science, such as productivity and collaboration.

team science | cumulative advantage

Contemporary science has undergone major changes in the last half century at all levels: institutional, intellectual, and social, as well as in its relationship with society at large. Science has been changing in response to increasingly complex problems of contemporary society and the inherently challenging nature of unresolved questions, with an expectation to serve as a major driver for economic growth. Consequently, the contemporary science community has adopted a problem-driven approach to knowledge production that often blurs the lines between pure and applied, and is more permeable around disciplinary borders, leading to cross-/multi/inter/transdisciplinarity (1). The major staple of this approach is team effort (2–5). The increased prominence of scientific teams has recently led to a formation of research area, “science of team science,” which is “centered on examination of the processes by which scientific teams organize, communicate, and conduct research” (6). If we want not only to understand contemporary science but also to create and promote viable science policies, we need to uncover principles that lead to the formation and subsequent evolution of scientific research teams.

Studies of collaboration in science, and coauthorship as its most visible form, have a long history (7–11). The collaborative mode of knowledge production is often perceived as being in contrast to the individualistic mode of the past centuries (12, 13). Previous studies have established that the fraction of coauthored papers has been growing with respect to single-authored papers (5), that in recent decades teams have been growing in size (14), and that interinstitution and international teams are becoming more prevalent (15, 16). In addition, high-impact research is increasingly attributed to large teams (5, 6), as is research that features more novel combination of ideas (17). The reasons for an increase in collaborative science have been variously explained

as due to the shifts in the types of problems studied (1) and the related need for access to more complex instruments and broader expertise (15, 18, 19).

A research team is a group of researchers collaborating to produce scientific results, which are primarily communicated in the form of research articles. Researchers who appear as authors on a research article represent a visible and easily quantifiable manifestation of a collaborative, team science effort. We refer to such a group of authors as an “article team.” In this study, we focus on one of the most fundamental aspects of team science: “article team size distribution” and its change/evolution over time. (In the rest of the article, we will refer to an article team simply as “the team.”) Many studies focused only on the mean or the median sizes of teams, implicitly assuming that the character of the distribution of team sizes does not change. Relatively few studies examined full team size distribution, albeit for rather limited datasets (10, 20, 21), with some of them noticing the changing character of this distribution (10). The goal of the current study is to present a more accurate characterization and go beyond empirical observations to provide a model of scientific research team formation and evolution that leads to the observed team size distributions.

Despite a large number of studies of coauthorship and scientific teams, there are few explanatory models. One such exception is the model by Guimerà et al. (2) of the self-assembly of teams, which is based on the role that newcomers and repeated collaborations play in the emergence of large connected communities and the success of team performance. Although their model features team size as a parameter, its values were not predicted by the model but were taken as input from the list of actual publications. The objective of the current study is to go beyond the internal composition of teams to explain the features of team size distribution and its change over the past half century.

Significance

Science is an activity with far-reaching implications for modern society. Understanding how the social organization of science and its fundamental unit, the research team, forms and evolves is therefore of critical significance. Previous studies uncovered important properties of the internal structure of teams, but little attention has been paid to their most basic property: size. This study fills this gap by presenting a model that successfully explains how team sizes in various fields have evolved over the past half century. This model is based on two principles: (i) smaller (core) teams form according to a Poisson process, and (ii) larger (extended) teams begin as core teams but consequently accumulate new members through the process of cumulative advantage based on productivity.

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

The author declares no conflict of interest.

This article is a PNAS Direct Submission.

¹E-mail: smilojev@indiana.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1309723111/-DCSupplemental.

Thus, the model we propose in this paper is complementary to the efforts by Guimerà et al. Our model is based on several simple principles that govern team formation and its evolution. The validity of the model is confirmed by constructing simulated team size distributions that closely match the empirical ones based on 150,000 articles published in the field of astronomy since the 1960s. We reveal the existence of two principal modes of knowledge production: one that forms small core teams based on a Poisson process, and the other that leads to large, extended teams that grow gradually on the principle of cumulative advantage.

Empirical Team Size Distributions

The significant change in the character of team size distribution is the key insight underlying the proposed model. Previous studies have shown a marked increase in the mean team size in recent decades, not only in astronomy (e.g., refs. 2 and 22), but in all scientific fields (5). Specifically, the average team size in astronomy grew from 1.5 in 1961–1965 to 6.7 in 2006–2010 (marked by arrows in Fig. 1, which shows, on a log-log scale, team size distributions in the field of astronomy in two time periods). However, Fig. 1 reveals even more: a recent distribution (2006–2010) is not just a scaled-up version of the 1961–1965 distribution shifted toward larger values; it has a profoundly different shape. Most notably, although in 1961–1965 the number of articles with more than five authors was falling precipitously, and no article featured more than eight authors, now there exists an extensive tail of large teams, extending to team sizes of several hundred authors. The tail closely follows the power-law distribution (red line in Fig. 1). The power-law tail is seen in recent team size distributions of other fields as well (25). In contrast, the “original” 1961–1965 distribution did not feature a power-law tail. Instead, most team sizes were in the vicinity of the mean value. The shape of this original distribution can instead be described with a simple Poisson distribution (blue curve in Fig. 1), an observation made in some previous works (10, 20).

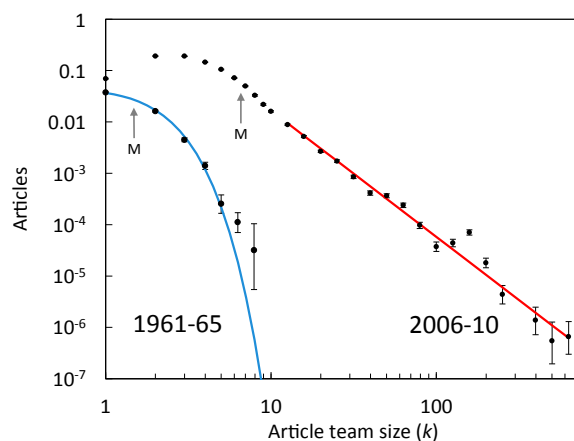


Fig. 1. Distribution of article team sizes in astronomy in two time periods separated by 45 y. The distribution from 1961 to 1965 is well described by a Poisson distribution (blue curve). This is in contrast to 2006–2010 distribution, which features an extensive power-law tail (red line). The arrows mark the mean values of each distribution. For $k > 10$ ($k > 5$ for 1961–1965), the data are binned in intervals of 0.1 decades, thus revealing the behavior far in the tail, where the frequency of articles of a given size is up to million times lower than in the peak. All distributions in this and subsequent figures are normalized to the 2006–2010 distribution in astronomy. Error bars in this and subsequent figures correspond to 1 SD. The full dataset consists of 154,221 articles published between 1961 and 2010 in four core astronomy journals (listed in *SI Materials and Methods*), which publish the majority of research in this field (23). Details on data collection are given elsewhere (24).

Note that the time when the distribution stopped being Poisson would differ from field to field.

We interpret the fact that the distribution of team sizes in astronomy in the 1960s is well described as a stochastic variable drawn from a Poisson distribution to mean that initially the production of a scientific paper used to be governed by a “Poisson process” (26, 27). This is an intuitively sound explanation because many real-world phenomena involving low rates arise from a Poisson process. Examples include pathogen counts (28), highway traffic statistics (29), and even sports scores (30). Team assembly can be viewed as a low-rate event, because its realization involves few authors out of a very large possible pool of researchers. Poisson rate (λ) can be interpreted as a characteristic number of authors that are necessary to carry out a study. The actual realization of the process will produce a range of team sizes, distributed according to a Poisson distribution with the mean being this characteristic number.

In contrast, the dynamics behind the power-law distribution that features in team sizes in recent times is fundamentally different from a simple Poisson process, and instead suggests the operation of a process of “cumulative advantage.” Cumulative advantage, also known as the Yule process, and as preferential attachment in the context of network science (31, 32), has been proposed as an explanation for the tails of collaborator and citation distributions (25, 32–38). Unlike the Poisson process, cumulative advantage is a dynamic process in which the properties of a system depend on its previous state. How did a distribution characterized by a Poisson function evolve into one that follows a power law? Does this evolution imply a change in the mode of the team assembly? Does a Poisson process still operate today? Fig. 1 shows that, for smaller team sizes ($k < 10$), the power law breaks down, forming instead a “hook.” This small- k behavior must not be neglected because the great majority of articles (90%) are still published in teams with fewer than 10 authors. The hook, peaking at teams with two or three authors, may represent a vestige of what was solely the Poisson distribution in the past. This simple assumption is challenged by the fact that no single Poisson distribution can adequately fit the small- k portion of the 2006–2010 team size distribution. Namely, the high ratio of two-author papers to single-author papers in the 2006–2010 distribution would require a Poisson distribution with $\lambda = 2P_{k=2}/P_{k=1} = 5.5$. Such distribution produces a peak at $k=5$, which is significantly offset compared with its actual position. Evidently, the full picture involves some additional elements.

In the following section, we present a model that combines the aforementioned processes and provides answers to the questions raised in this section, demonstrating that knowledge production occurs in two principal modes.

Model of Team Formation and Evolution

We next lay out a relatively simple model that incorporates principles of team formation and its evolution. We produce simulated team size distributions based on the model and validate them by testing how well they “predict” empirical distributions in the field of astronomy. This model is universally applicable to other fields, as will be discussed later.

The model consists of authors who write papers over time. Each paper has a “lead” author who is responsible for putting together a team and producing a paper. Each lead author is associated with two types of teams: “core” and “extended.” Core teams consist of the lead author and coauthors. Their size is drawn from a Poisson distribution with some rate λ . If the drawing yields the number 1, the core team consists of the lead author alone. We allow λ , the characteristic size of core teams, to grow with time. Existing authors, when they publish again, retain their original core teams. The probability of publishing by an author who has published previously is 0.8. Unlike core teams, extended teams evolve dynamically. Initially, the extended team

Table 1. Characteristics of different fields obtained from analytical decomposition

Field	Articles (2006–2010)	λ_{P1}	λ_{P2}	α_{PL}	f_{P1}	f_{P2}	f_{PL}	μ_P	μ_{PL}	μ_{all}
Astronomy	31,473	3.25	0.67	2.8	0.52	0.11	0.37	3.21	11.20	6.14
Ecology	5,420	3.23	0.83	3.8	0.62	0.13	0.25	3.20	4.58	3.54
Mathematics	3,244	0.87	0.75	13.4	0.57	0.33	0.09	1.84	2.87	1.93
Social psychology	4,122	2.24	1.58	4.5	0.46	0.36	0.18	2.72	3.89	2.93
Literature	725	0.06	0.02	5.0	0.99	0.00	0.01	1.03	3.75	1.05
arXiv	235,414	1.80	4.93	2.6	0.72	0.05	0.23	2.38	6.56	3.36

The meaning of the columns is given in the legend of Fig. 6.

authorship, as derived by the model (Fig. 3). By integrating these components, we find that currently 57% of articles belong to F_{P1} and can therefore be attributed to standard core teams. Another 12% are due to “core +1” teams (F_{P2}), whereas the remaining 31% of articles are fit by the truncated power-law component (F_{PL}) and can therefore be interpreted as originating from extended teams.

The principles that underlie the proposed model are universal and not field dependent. Only the parameters that specify the rate of growth or the relative strength of the processes will differ from field to field. Consequently, the analytical decomposition given by Eq. 1 can be applied to other fields. Fig. 5 shows the best-fitting functions (Eq. 1) to the empirical team size distributions in the following fields: mathematics, ecology, social psychology, literature, and for articles from arXiv, all for the current period (2006–2010). Core journals used for these fields are listed in *SI Materials and Methods*. All of the distributions are well described by our model-based functional decomposition. Parameters for the fit and contributions of different authorship modes are given in Table 1. There is much variety. In literature, the standard core team mode accounts for nearly the entire output (99%) with very small teams. Mathematics also features relatively small teams and a steep decline of larger teams. Nevertheless, the functional decomposition implies that 9% of articles are produced in the extended-team mode (see also Fig. S3), but these teams are still not much larger than core teams (2.9 vs. 1.8 members on average). Mathematics and social psychology feature the largest share of “core +1” teams. Team size distributions for ecology and social psychology both have more prominent power-law tails than mathematics ($\alpha \sim 4$), but they are not yet as extensive as in astronomy ($\alpha \sim 3$). Both fields feature a hook at low k similar to that of astronomy. Finally, articles from arXiv (mostly belonging to the field of physics) have a power-law slope very similar to that of astronomy.

Application of Analytical Decomposition for Describing Trends in Team Evolution

Analytical decomposition, introduced in the previous section, allows us to empirically derive the contribution of different modes of authorship over time and to explore the characteristics of teams as they evolve. We now fit Eq. 1 to article teams in astronomy for all 5-y time periods, from 1961 to 2010. Fig. 6, *Left*, shows the change in the best-fit Poisson rates of both types of core teams as well as the evolution of the slope of the power-law component. As previously suggested, the Poisson rate of core teams has gradually increased from close to zero in the early 1960s to a little over three recently. However, the slope of the power-law component has gradually been flattening, from $\alpha=6$ to $\alpha=3$, i.e., the power-law component has been gaining in prominence.

Fig. 6, *Center*, shows the relative contributions of the three modes of authorship in astronomy over the time period of 50 y, obtained by integrating the best-fit functional components. Remarkably, the contributions have remained relatively stable, with articles in the power-law component (i.e., articles produced by extended teams) making $\sim 30\%$. This stability in the fraction of power-law articles is directly connected to the fixed propensity of authors to write articles with extended teams, as indicated in the model simulation. In all time periods, most papers ($\sim 60\%$) have been published by standard core teams (the Poisson component). Core teams with an extra member seem to appear in the early 1970s, but their contribution has remained at around 10%.

As pointed out earlier, many studies have emphasized the impressive growth of “mean” team sizes. We can now explore this trend in the light of the various authorship modes. In Fig. 6, *Right*, we show the change in the mean size of all teams, and separately of core teams (standard and “core +1” teams together) and of power-law (extended) teams. In the early 1960s, both the core and the extended teams were relatively small (1.1

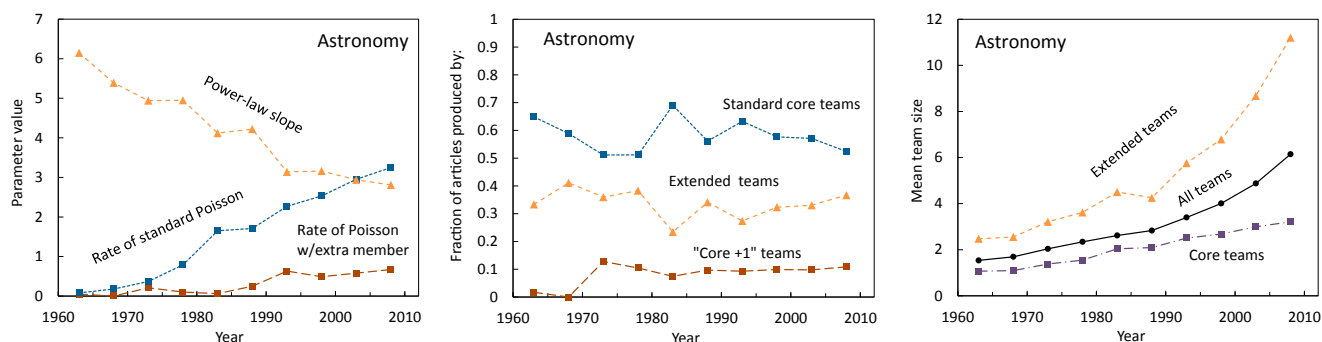


Fig. 6. Trends in team evolution in astronomy from 1961 to 2010. *(Left)* Fifty-year trend of parameters characterizing the three components of the distribution, derived from a functional fit (Eq. 1). The characteristic size (i.e., Poisson rate) of standard core teams has (λ_{P1}) been rising throughout this period, whereas that of “core +1” teams (λ_{P2}) has remained constant in the last two decades. The power-law slope (α_{PL}) has been getting shallower, i.e., the significance of the power-law component has been increasing. *(Center)* Fraction of articles produced by different modes of authorship (team types): standard core (f_{P1}), “core +1” (f_{P2}), and extended (f_{PE}). *(Right)* Trends in the mean team size, overall (μ_{all}) and by team type (both types of core teams, μ_P , and extended teams, μ_{PE}). The increase in the overall mean team size in astronomy is primarily the result of the rapid growth of power-law (extended) teams.

and 2.5 members, respectively). Subsequently, the mean size of core teams has increased linearly to 3.2 members. However, the mean size of extended teams has grown exponentially, and most recently averages 11.2 members. The exponential increase in the size of extended teams is affecting the overall mean, despite the fact that the extended teams represent the minority mode of authorship. Although the growth of core teams is more modest, it nevertheless indicates that the level of collaboration, as measured by article team size, increases for this traditional mode of producing knowledge as well. Whether this increase is a reflection of a real change in the level of collaborative work or simply a change in the threshold for a contributor to be considered a coauthor is beyond the scope of this work.

In a similar fashion, we explored the evolution of fit parameters, mode contributions, and team sizes for mathematics and ecology (Figs. S3 and S4). Mathematics features a small extended-team component (10%) that emerged in the mid-1980s. Extended teams in mathematics are still only slightly larger in size than the core teams. The share of “core +1” teams is increasing. The mean size of all core teams has increased, albeit moderately (from 1.2 to 1.8 members). In ecology, the overall increase in mean team size mostly reflects the increase of the characteristic size of standard core teams in the 1980s. The observed increase of the share of extended teams appears to come at the expense of standard core teams.

Implications and Conclusions

The model proposed in this paper successfully explains the evolution of the sizes of scientific teams as manifested in author

lists of research articles. It demonstrates that team formation is a multimodal process. Primary mode leads to relatively small core teams, the size of which may represent the typical number of researchers required to produce a research paper. The secondary mode results in teams that expand in size, and which are presumably used to carry out research that requires expertise or resources from outside of the core team. These two modes are responsible for producing the hook and the power law-tail in team size distribution, respectively.

This two-mode character may not be exclusive to team sizes. Interestingly, a similarly shaped distribution consisting of a hook and a power-law tail is characteristic of another bibliometric distribution, that of the number of citations that an article receives. Recently, a model was proposed that successfully explained this distribution (33) by proposing the existence of two modes of citation, direct and indirect, where the latter is subject to cumulative advantage.

Understanding the distribution of the number of coauthors in a publication is of fundamental importance, as it is one of the most basic distributions that underpin our notions of scientific collaboration and the concept of “team science.” The principles of team formation and evolution laid out in this work have the potential to illuminate many questions in the study of scientific collaboration and communication, and may have broader implications for research evaluation.

ACKNOWLEDGMENTS. I thank the two anonymous reviewers for their constructive comments and Colleen Martin and John McCurley for copyediting.

- Gibbons M, et al. (1994) *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies* (Sage, London).
- Guimerà R, Uzzi B, Spiro J, Amaral LA (2005) Team assembly mechanisms determine collaboration network structure and team performance. *Science* 308(5722):697–702.
- Jones BF, Wuchty S, Uzzi B (2008) Multi-university research teams: Shifting impact, geography, and stratification in science. *Science* 322(5905):1259–1262.
- Newman MEJ (2004) *Who is the Best Connected Scientist? A Study of Scientific Co-authorship Networks*. *Complex Networks*, eds Ben-Naim E, Frauenfelder H, Toroczkai Z (Springer, Berlin), pp 337–370.
- Wuchty S, Jones BF, Uzzi B (2007) The increasing dominance of teams in production of knowledge. *Science* 316(5827):1036–1039.
- Börner K, et al. (2010) A multi-level systems perspective for the science of team science. *Sci Transl Med* 2(49):49cm24.
- Price DJS (1963) *Little Science, Big Science* (Columbia Univ Press, New York).
- DD B (1978) Possible relationships between the history and sociology of science. *Social Inq* 48(3-4):140–161.
- Babchuk N, Keith B, Peters G (1999) Collaboration in sociology and other scientific disciplines: A comparative trend analysis of scholarship in the social, physical, and mathematical sciences. *Am Sociol* 30(3):5–21.
- Glänzel W (2002) Coauthorship patterns and trends in the sciences (1980–1998): A bibliometric study with implications for database indexing and search strategies. *Libr Trends* 50(3):461–473.
- Kretschmer H (1997) Patterns of behaviour in coauthorship networks of invisible colleges. *Scientometrics* 40(3):579–591.
- Cronin B (2001) Hyperauthorship: A postmodern perversion of evidence of a structural shift in scholarly communication practices? *J Am Soc Inf Sci Technol* 52(7):558–569.
- Cronin B, Shaw D, La Barre K (2003) A cast of thousands: Co-authorship and sub-authorship collaboration in the twentieth century as manifested in the scholarly literature of psychology and philosophy. *J Am Soc Inf Sci Technol* 54(9):855–871.
- Bordons M, Gomez I (2000) Collaboration networks in science. *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*, eds Cronin B, Atkins HB (Information Today, Medford, NJ), pp 197–213.
- Shrum W, Genuth J, Chompalov I (2007) *Structures of Scientific Collaboration* (MIT, Cambridge, MA).
- Wagner CS (2008) *The New Invisible College: Science for Development* (Brookings Institution, Washington, DC).
- Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. *Science* 342(6157):468–472.
- Hagstrom WO (1965) *The Scientific Community* (Basic Books, New York).
- Melin G (2000) Pragmatism and self-organization: Research collaboration on the individual level. *Res Policy* 29(1):31–40.
- Price DJ, Beaver DD (1966) Collaboration in an invisible college. *Am Psychol* 21(11):1011–1018.
- Epstein RJ (1993) Six authors in search of a citation: Villains or victims of the Vancouver convention? *BMJ* 306(6880):765–767.
- Fernández JA (1998) The transition from an individual science to a collective one: The case of astronomy. *Scientometrics* 42(1):61–74.
- Henneken EA, et al. (2007) E-print journals and journal articles in astronomy: A productive co-existence. *Learn Publ* 20(1):16–22.
- Milojević S (2012) How are academic age, productivity and collaboration related to citing behavior of researchers? *PLoS One* 7(11):e49176.
- Milojević S (2010) Modes of collaboration in modern science—beyond power laws and preferential attachment. *J Am Soc Inf Sci Technol* 61(7):1410–1423.
- Kingman JFC (1993) *Poisson Processes* (Oxford Univ Press, Oxford).
- Ross SM (1995) *Stochastic Processes* (Wiley, New York), 2nd Ed.
- Feller W (1968) *An Introduction to Probability Theory and Its Applications* (Wiley, New York), 3rd Ed.
- Gerlough DL, Andre S (1955) *Use of Poisson Distribution in Highway Traffic. The Probability Theory Applied to Distribution of Vehicles on Two-Lane Highways* (Eno Foundation for Highway Traffic Control, Saugatuck, CT).
- Karlis D, Ntzoufras I (2003) Analysis of sports data by using bivariate Poisson models. *Statistician* 52(3):381–393.
- Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. *Contemp Phys* 46(5):323–351.
- Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. *SIAM Rev Soc Ind Appl Math* 51(4):661–703.
- Peterson GJ, Pressé S, Dill KA (2010) Nonuniversal power law scaling in the probability distribution of scientific citations. *Proc Natl Acad Sci USA* 107(37):16023–16027.
- Barabási A-L, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509–512.
- Barabási A-L, et al. (2002) Evolution of the social network of scientific collaborations. *Physica A* 311(3-4):590–614.
- Newman MEJ (2001) Clustering and preferential attachment in growing networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 64(2 Pt 2):025102.
- Newman MEJ (2001) The structure of scientific collaboration networks. *Proc Natl Acad Sci USA* 98(2):404–409.
- Börner K, Maru JT, Goldstone RL (2004) The simultaneous evolution of author and paper networks. *Proc Natl Acad Sci USA* 101(Suppl 1):S266–S273.
- Abdo AA, et al. (2010) Fermi large area telescope first source catalog. *Astrophys J Suppl Ser* 188(2):405–436.
- Moody J (2004) The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *Am Sociol Rev* 69(2):213–238.

Supporting Information

Milojević 10.1073/pnas.1309723111

SI Materials and Methods

The following is an alphabetical list of the core journals that were used in the analysis:

Astronomy (four journals): *Astronomical Journal*, *Astronomy & Astrophysics*, *Astrophysical Journal*, and *Monthly Notices of the Royal Astronomical Society*.

Ecology (five journals): *Ecology*, *Journal of Animal Ecology*, *Journal of Ecology*, *Oecologia*, and *Oikos*.

Literature (six journals): *English Literary History*, *English Literature in Transition (1880–1920)*, *Modern Fiction Studies*, *New Literary History*, *Publications of the Modern Language Association*, and *Twentieth Century Literature*.

Mathematics (eight journals): *Acta Mathematica*, *American Journal of Mathematics*, *Annals of Mathematics*, *Inventiones Mathematicae*, *Journal of Functional Analysis*, *Journal of the American Mathematical Society*, *Mathematics of Computation*, and *Proceedings of the London Mathematical Society*.

Social psychology (nine journals): *British Journal of Social Psychology*, *European Journal of Social Psychology*, *Journal of Applied Social Psychology*, *Journal of Experimental Social Psychology*, *Journal of Personality and Social Psychology*, *Journal of Social Psychology*, *Personality and Social Psychology Bulletin*, *Social Behavior and Personality*, and *Social Psychology Quarterly*.

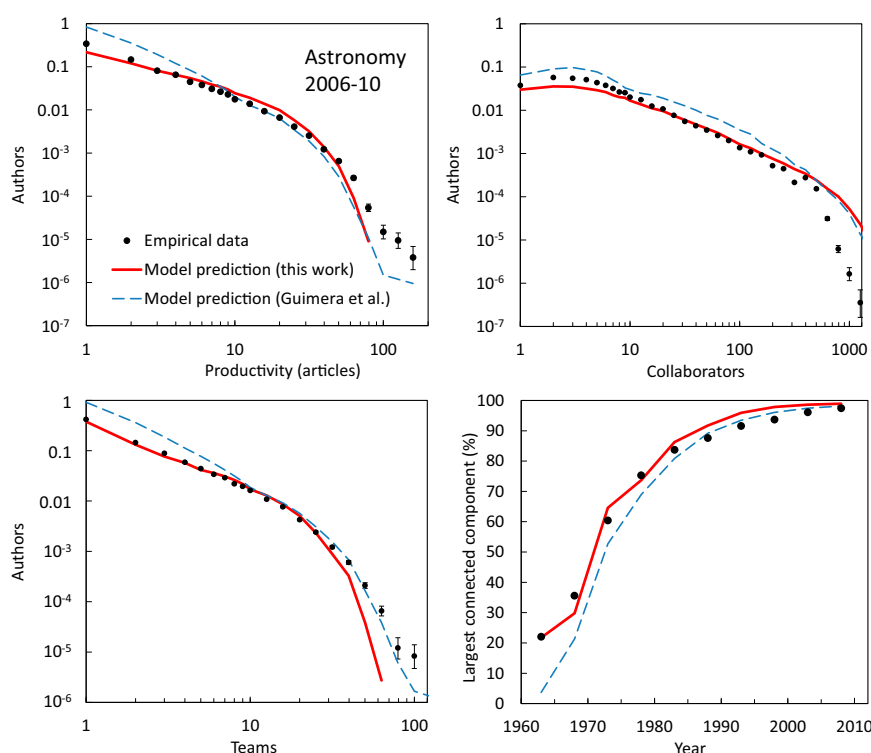


Fig. S1. Comparison of empirical data and model predictions for the distributions of author productivity (*Upper Left*), number of collaborators (*Upper Right*), and teams per author (*Lower Left*), as well as the trend of the relative size of the largest connected component (*Lower Right*), for astronomy (2006–2010). Productivity (number of articles per author) is determined based on all articles regardless of author role (lead author or coauthor). Collaborators are defined as all coauthors of a given author from all articles published during this period. The number of teams with which a given author is involved is determined as the number of different lead authors in all of the papers in which the given author is a coauthor. The largest connected component is the fraction of all authors that are linked by coauthorship in this time period. Values above 50% signal the emergence of a giant component. Model predictions are given for the model featured in this work (red solid lines) and, as a comparison, based on the principles of team assembly laid out in ref. 1 (blue dashed lines). The model by Guimerà et al. requires explicit input on team size for each article, which we take from our model. Then, for the Guimerà et al. predictions, we choose team members by assuming probability p that the team member is an author already present in the network (incumbent), and probability q that the incumbent is already a collaborator of some other team member. We use $p = 0.78$ and $q = 0.82$, as appropriate for astronomy (ref. 1, figure 4J). Also, to recreate the Guimerà et al. process fully, we implement removal of authors who have been inactive for longer than 5,000 time steps (ref. 1, supplemental information). Both the current and Guimerà et al. methods for team buildup provide good general description of author-centric distributions and of the evolution of coauthorship network topology.

1. Guimerà R, Uzzi B, Spiro J, Amaral LA (2005) Team assembly mechanisms determine collaboration network structure and team performance. *Science* 308(5722):697–702.

