Task specialization and its effects on research careers

Nicolas Robinson-Garcia[1]✉, Rodrigo Costas[2,3], Cassidy R. Sugimoto[4], Vincent Larivière[5], and Gabriela F. Nane[1]

[1]Delft Institute of Applied Mathematics (DIAM), TU Delft, 2628 Delft, Netherlands
[2]Centre for Science and Technology Studies, Leiden University, 2333 AL Leiden, The Netherlands
[3]Centre for Research on Evaluation, Science and Technology (CREST), Stellenbosch University, Stellenbosch, South Africa
[4]School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, IN 47408
[5]École de bibliothéconomie et des sciences de l'information, Université de Montréal, H3T 1N8 Montreal, Canada

We model a set of 70,694 publications and 347,136 distinct authors using Bayesian networks to predict scientists' specific contributions on each of their publications. We predict the contributions of 222,925 authors in 6,236,239 publications, and apply an archetypal analysis to profile scientists by career stage. We divide scientific careers into four stages: junior, early-career, mid-career and late-career. Three scientific archetypes are found throughout the four career stages: leader, specialized, and supporting. All three archetypes are encountered for the early- and mid-career stages, whereas for junior and late-career stages only two archetypes are found. Scientists assigned to the leader and specialized archetypes tend to have longer careers than researchers who belong to the supporting archetype. There is consistent gender bias at all stages: the majority of male scientists belong to the leader archetype, while the larger proportion of women belong to the specialized archetype, especially for early and mid-career researchers.

science of science | research careers | scientometrics | gender in science | research evaluation

Correspondence: ***n.robinsongarcia@tudelft.nl***

Introduction

The assessment of scientific careers has been under scrutiny for some time (1–3). Successful careers are built on concepts such as leadership (4), productivity (1, 5), and impact (6, 7). However, evidence suggests that the design of a unique career path built on individualistic success may hamper the way in which science is actually produced (8). Collaboration has become essential and ubiquitous in science (8–10); however, the increase in team size may come at a cost for those who are not in leading roles (8). Recent evidence shows an increasing need for a larger and more stratified scientific workforce (11–14) which necessarily involves a reconceptualization of research careers and considering a breadth of profiles for which specific paths should be considered. The overreliance on past success (15, 16) may both reduce the scientific careers of team players (8) and introduce gender biases (17–19), discouraging women to pursue a career in academia (20, 21). This heterogeneity in scientists' profiles realizes the need for distribution of labor (12). However, there is still a lack of understanding of how research profiles differ from each other, and how they are associated with career stages (22).

The goal of this study is to analyze the relation between task specialization and career length of scientists. Do specific profiles of scientists have shorter research careers than others? How do profiles relate to gender? Are these differences also reflected in productivity and citations? To answer those questions, we develop a Bayesian network–that is, a probabilistic graphical model–to predict the specific contributions scientists made to each of their publications throughout their career. We then profile researchers based on their contributions and explore how those profiles evolve throughout their careers. We investigate how profiles at each career stage affect career length, with a particular focus on the relationship with the perceived gender of the scientist. Finally, we examine the relationship between profiles and bibliometric characteristics, such as research production and scientific impact.

Our seed dataset contains a total of 70,694 papers authored by 347,136 scientists from the Medical and Life Sciences. Author names are disambiguated using a rule-based scoring algorithm (23). Each author has also been linked to their bibliometric data from Web of Science. We restrict our dataset to the Medical and Life Sciences to make it more homogeneous and avoid disciplinary differences in task distribution. We assign papers to fields by identifying the journal to which each of the references of the publications in our dataset belong. Then, we assign to each publication, the field from which most of its references come. Finally, we only include those which are assigned to the Medical and Life Sciences fields. Further details are provided in the Materials and Methods section.

We then build a probabilistic model to predict authors' contribution to publications, based on a set of bibliometric variables. This model allows us to extend our analysis from the initial dataset to the complete publication history of these authors. We reconstruct the publication history of 222,925 authors from our original dataset and predict, for each author, the probability of conducting a given contribution on each of their publications. Based on the new dataset of predicted probabilities of contributorship, we divide scientists' careers into four stages and conduct an archetypal analysis (24) by stage. This allows us to identify career paths and discuss differences in scientific profiles by stage, scientific paths, and gender.

Results

Contribution statements and predicting variables. Five types of contributions are identified in the contribution dataset: wrote the paper (WR), conceived and designed the experiments (CE), performed the experiments (PE), analyzed the data (AD), and contributed reagents/materials/analysis tools (CT). Furthermore, the number of contributions (NC) is also considered. These contributions are related to author position (8, 10, 25), with first and last positions in author order reflecting leadership (26), as per the recommendations of the (27). Figure 1 relates career stage and author order with contribution role. We define four career stages: junior (< 5 years since first publication), early-career (≥ 5 and < 15 years since first publication), mid-career (≥ 15 and < 30 years since first publication) and late-career (≥ 30 years since first publication).

The distribution of contribution roles by career stage shows that earlier stages are more often associated with performing experiments and analyzing data, and that this contribution decreases as individuals become more senior. Writing the manuscript and contributing reagents and tools increase over time, with a decline in the late-career stage. Conceiving and designing the experiments demonstrates a modal shape, where early-career and mid-career stages are the ones in which these tasks are more prominent. In terms of labor distribution, first authors are heavily associated with all contributions, with the exception of contributing tools, reagents, data, and other materials. Middle authors are less involved in writing tasks or in the design and conception of experiments but are associated with contributing resources to a much greater extent. Lastly, authors contribute mostly to the design and conception of experiments as well as to writing tasks, and to a lesser extent to the performance of experiments.

Bibliometric indicators are employed as predictors of contributorship. Two types of bibliometric variables are included: paper-level and author-level. Paper-level variables are document type (DT), number of authors (AU), number of countries (CO), and institutions (IN) to which authors of the paper are affiliated. Author-level variables include their position in the authors' list (PO), number of years since they published their first publication (YE) and the average number of publications per year (PU).

Figure 2A depicts the Spearman rank correlation matrix of the contributorship and bibliometric data, while Figure 2B illustrates the Bayesian network used for predicting the contribution of a researcher for a given publication. The highest correlations within types of contributorship are between writing the manuscript and conceiving and designing the experiments (0.51), while the rest of contributorship variables exhibit low correlations. In the case of bibliometric variables, there is a moderate positive correlation between number of countries and institutions (0.57), author position and number of authors (0.54), and number of authors and number of institutions (0.51). A positive monotone relation between the number of contributions and either writing the manuscript (0.69), conceiving the experiments (0.66) or analyzing the data (0.64) is observed. Weak monotone relationships are suggested by correlations between contributorship and bibliometric variables are observed. A negative correlation of -0.38 is observed between performing the experiments and position in authors list and years since publication. Weak to moderate negative correlations are observed between contributorship variables and the number of countries and institutions, author's position, and number of authors of a publication.

Bayesian network model for predicting contributorship. We model our dataset using a Bayesian network (BN) to be able to predict contribution roles of scientists for their publications. The aim here is to expand our original dataset to the complete publication history of the 347,136 researchers from the Medical and Life Sciences who had published at least one paper in our PLOS seed dataset. A BN is a probabilistic graphical tool used to model multivariate data (28). The variables are denoted as nodes in the network, whereas the arcs denote influences between variables, typically quantified as dependencies.

Figure 2B shows the structure of the obtained BN. Five types of contributions along with the number of contributions (in green) of scientists are predicted using the seven bibliometric variables (in blue). The structure of the BN has been obtained by using a hybrid data-learning algorithm called Max-Min Hill Climbing (MMHC) (29), along with the constraint that bibliometric variables are influencing contributorship variables. That is, if an arc between bibliometric and contributorship variables is present in the structure, then it should be directed to the contributorship variable. Furthermore, the structure of the network has been tested for robustness. The strength of the arcs, i.e., relationships between variables, has been investigated using the bootstrap procedure, with 50 repetitions. Only the arcs that were present in 80% of the repetitions have been considered and are depicted in Figure 2B.

We evaluate the predictive power of the obtained BN using k-fold cross-validation. That is, the data has been repeatedly divided in 10 random folds, of which 9 have been used to learn the BN structure using the MMHC algorithm together with the aforementioned constrained. The contributions were then predicted for the remaining fold. The procedure has been repeated for each of the 10 folds and results on the prediction errors are reported in the Data and Methods section. The predictive performance of the BN has been shown to be extremely good, with an average classification error rate of between 6-8% for all contributorships. The BN is used to predict the contributions for the complete publication history of a subset of 222,925 scientists who have published in PLOS journals, for a total of 6,236,239 publications. Each contribution is predicted as the probability that an author has performed a given contribution on a publication. We further investigate the distributions of the predicted contributorships. When distinguishing by career stage (Figure 3), the densities clearly depict differences in contributorships. Performing the experiments is the most discriminative contributorship type, with junior scientists more likely be associated with this contribution. The more scientists advance in their career, the less likely that they will perform the experiments. Albeit less dramatic, the same discriminative pattern can be observed for analyzing the data and for the total number of contributorships, with decreasing association by age. Inversely, the

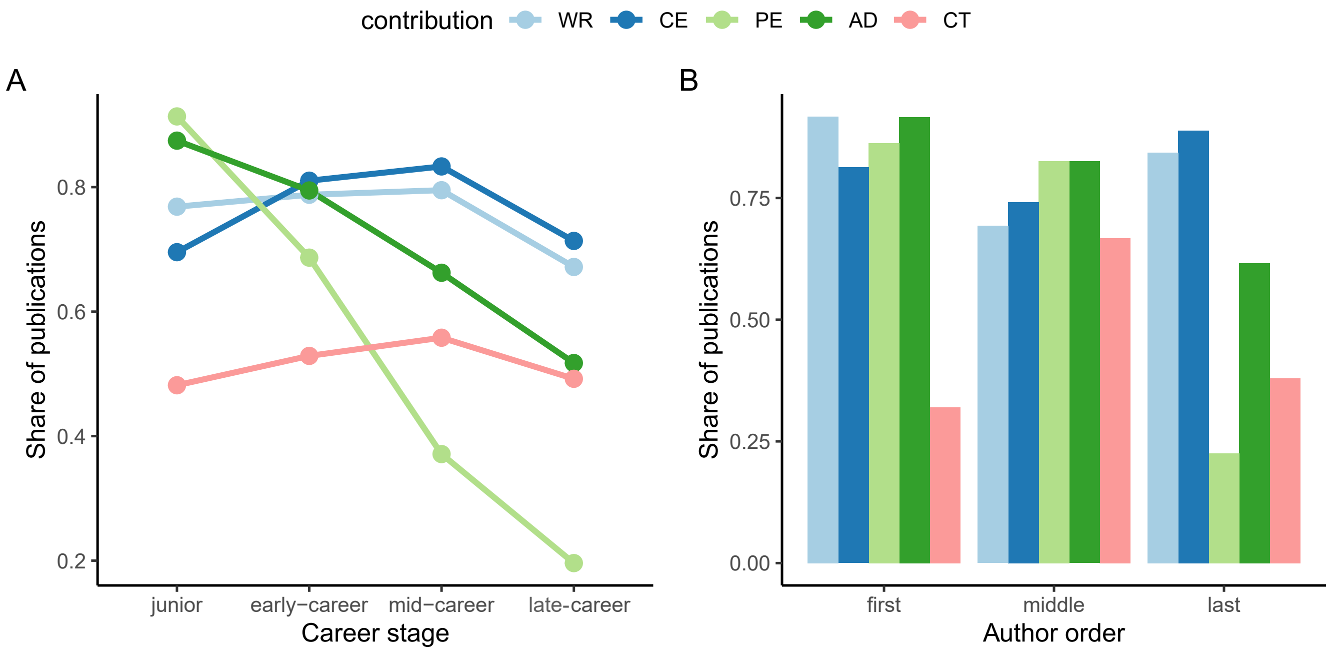
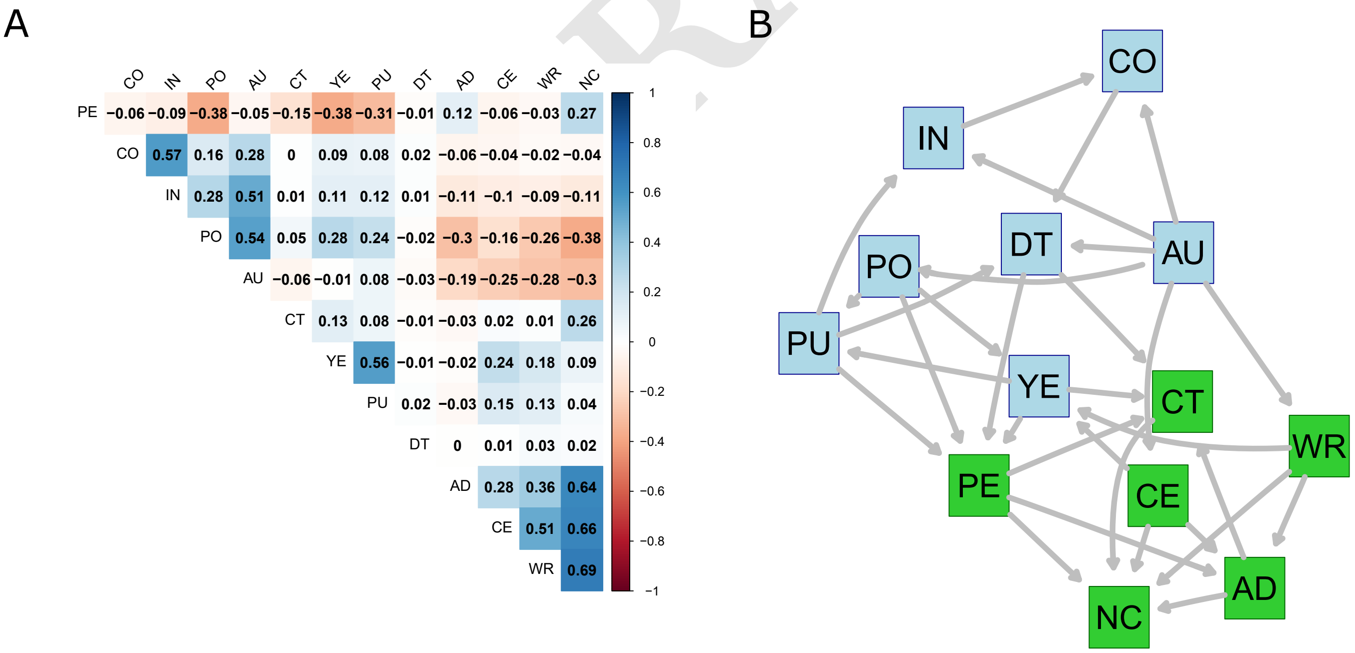


Fig. 1. Distribution of contributions by career stage **(A)** and author order **(B)**. In B only publications with at least 3 authors are included. **Career stages:** junior stage (< 5 years since first publication); early-career stage (≥ 5 and < 15 since first publication); mid-career stage (≥ 15 and < 30 years since first publication); and full career stage (≥ 30 years since first publication). WR (wrote the paper); AD (analyzed the data); CE (conceived and designed the experiments); CT (contributed reagents/materials/analysis tools); PE (performed the experiments); NC (number of contributions).



	CO	IN	PO	AU	CT	YE	PU	DT	AD	CE	WR	NC
PE	−0.06	−0.09	−0.38	−0.05	−0.15	−0.38	−0.31	−0.01	0.12	−0.06	−0.03	0.27
CO		0.57	0.16	0.28	0	0.09	0.08	0.02	−0.06	−0.04	−0.02	−0.04
IN			0.28	0.51	0.01	0.11	0.12	0.01	−0.11	−0.1	−0.09	−0.11
PO				0.54	0.05	0.28	0.24	−0.02	−0.3	−0.16	−0.26	−0.38
AU					−0.06	−0.01	0.08	−0.03	−0.19	−0.25	−0.28	−0.3
CT						0.13	0.08	−0.01	−0.03	0.02	0.01	0.26
YE							0.56	−0.01	−0.02	0.24	0.18	0.09
PU								0.02	−0.03	0.15	0.13	0.04
DT									0	0.01	0.03	0.02
AD										0.28	0.36	0.64
CE											0.51	0.66
WR												0.69

Fig. 2. Spearman correlation matrix of contributorship and bibliometric variables **(A)** and the Bayesian network used for predicting contributorship **(B)**. Contribution variables are in green, bibliometric variables are in blue. *Bibliometric variables*: PO (author's position); AU (number of authors); DT (document type); CO (number of countries); IN (number of institutions); YE (years since first publication); PU (average number of publications). *Contribution variables*: WR (wrote the paper); AD (analyzed the data); CE (conceived and designed the experiments); CT (contributed reagents/materials/analysis tools); PE (performed the experiments); NC (number of contributions).

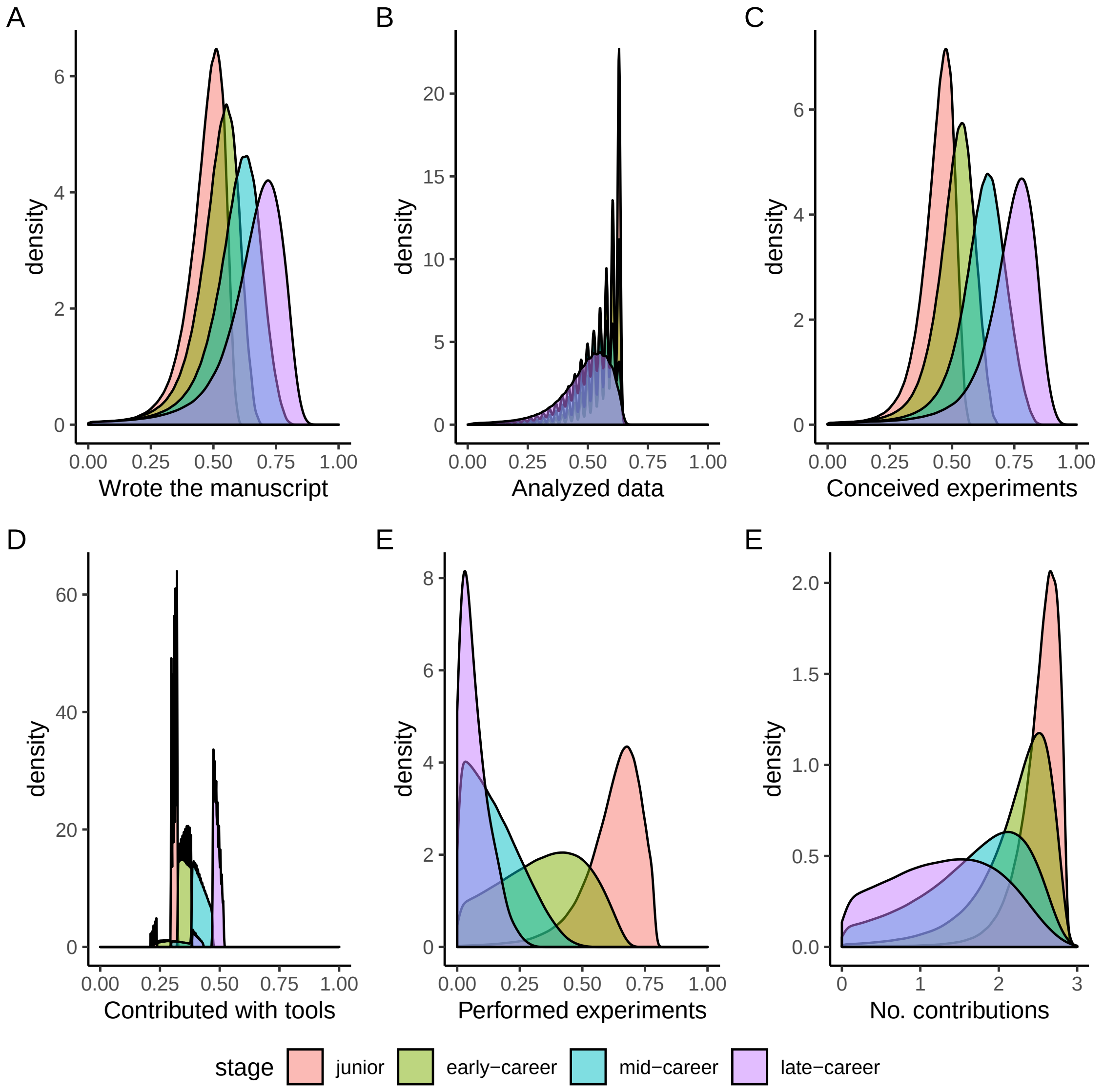


Fig. 3. Probability density functions of contribution roles predicted using the Bayesian Network model. Distributions are aggregated by career stage.

contribution roles of wrote the manuscript, conceived the experiments, and contributed with tools are more likely for advanced career stages.

Profiling scientists using archetypal analysis. We aggregate the predicted contributorships at the individual level and by career stage to profile scientists based on their contributorship patterns. To avoid the effect of contributorship outliers, we aggregate researchers' contributorships by choosing the median predicted contributorship of publications for each career stage. We perform a robust archetypal analysis to identify types of scientists based on their contributorships (30). Archetypes accentuate distinct features of scientists based on contribution data. Robust archetypal analysis identifies "prototypical types" of the multivariate aggregated contributorship dataset, correcting for outlier effects in the data. Each of these "prototypical types" or archetypes is represented as a convex combination of researchers in the aggregated contributorship dataset and, in turn, each researcher is well described by a convex combination of these archetypes.

We consider archetypes of scientists at each career stage. A residual sum of squares (RSS) analysis for different archetypes reveals that using two archetypes for the junior and late-career stages, and three for early-career and mid-career stages results in significantly smaller RSS. Figure **??** reveals the screeplots of RSS per career stage, where the elbow criterion supports the choice of number of archetypes per career stage. The influence of contributorships within each archetype is captured by corresponding coefficient values. Coefficients of each archetype (Leader, Specialized and Supporting) per career stage are presented in Figure 4. Low values indicate low prevalence of corresponding type of contributorship, whereas high values indicate a high contribution to the archetype.

A first notable observation is that differences in contributions are remarkably small for certain archetypes throughout career stages. Given that the archetypes at each stage have common characteristics, we maintain the same profile naming across stages. Three archetypes are identified. The Leader is charac-

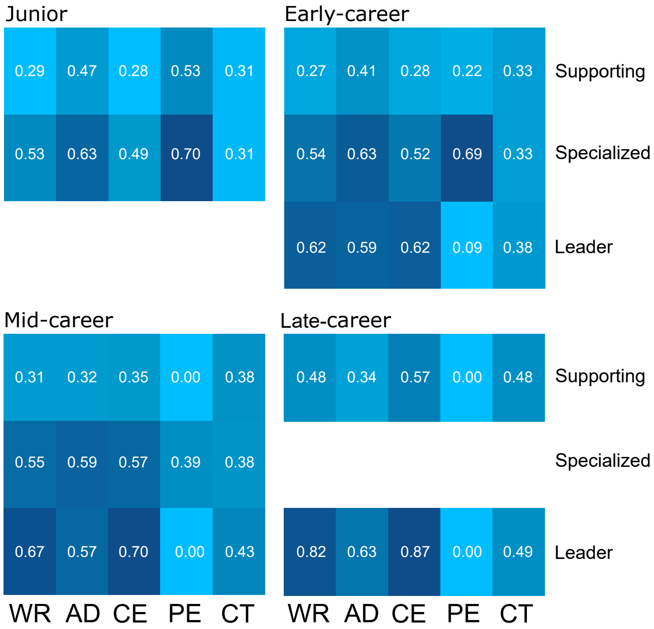


Fig. 4. Coefficient values of contributorships by archetype, per career stage. Two archetypes are identified in the junior stage (*Specialized* and *Supporting*), three have been identified for the early- and mid-career (*Leader*, *Specialized* and *Supporting*) and two have been identified for the late-career stage (*Leader* and *Supporting*).

terized by high coefficient values for all contributions, except for PE, indicating a high prevalence of each contribution role, and especially on WR and CE. The Specialized archetype is characterized by high coefficient values for PE and AD. A trend analysis for this archetype indicates a shift between PE and AD contributions. The third archetype is referred to as the Supporting, and is characterized by generally low values for all contributorships. This is the least discriminatory archetype.

At the junior stage, we observe two archetypes: Specialized and Supporting. Both are characterized by scientists reporting more than two contributions per paper. For the Specialized archetype, the most prevalent roles are on PE and AD, although they show higher coefficients than Supporting for all contributorships except CT (with a marginal difference). At the early-career stage, three archetypes are obtained, with a clear difference on PE between Leader and Specialized. These three archetypes are maintained during the mid-career stage, with the most notable difference being the shift between AD and PE for the Specialized, that now exhibits a higher probability of conducting the former than the latter. In the late-career stage, the Specialized archetype is no longer identified, and again two archetypes emerge. Both archetypes show low probabilities on PE, while the Leader is characterized by a higher probability on WR and CE. Overall, the archetypal analysis shows that the predictions obtained from the BN can accurately capture the diversity of archetypes of scientists and are sufficiently discriminating.

Career paths, productivity and citation impact. Similarities between the archetypes are identified at each career stage, demonstrating the stability of the classification by scientific age (Figure 4). In turn, each scientist can be represented as a weighted combination of the archetypes. For a given scientist, the weights, or α scores, corresponding to each archetype determine the researchers' assignment to one of the two or three archetypes. Here, we assign researchers to archetypes based on the highest weight. The assignment can be done for each career stage, which naturally leads to a career path.

Figure 5A presents the assignment of researchers to the archetypes and their evolution over the four career stages, using the maximum coefficients and the median aggregation method. However, we observe some patterns by archetype. Out of the 222,295 scientists included in the dataset, 27,714 reached the late-career stage. We observe that there is little attrition, regardless of the archetype to which scientists belong, between the junior and early-career stage (93% for junior Specialized and 83% for Supporting authors). At the early-career stage, when the Leader archetype emerges, the advantage of those exhibiting a Leader profile becomes evident: 84% of scientists who belong to the Leader archetype in their early-career reach the next career stage, while 30% and 16% of Specialized and Supporting scientists progress to mid-career stage respectively. The cost is even higher from mid-career to late-career, with 37% of Leader profile scientists, and only 1% and 2% of Specialized and Supporting authors reaching the last career stage.

Furthermore, 98% of scientists reaching the late-career stage exhibited a Specialized archetype in their junior stage, and 67% of those reaching this last career stage have consistently displayed a Leader profile in early- and mid-career stages. Shifts across archetypes appear more likely at earlier career stages, as well as from the Leader archetype to the other two archetypes (but not vice versa). Even though most of the scientists reaching the late-career stage belong to the Leader archetype in their mid-career stage, 66% of late-career researchers are in a Supporting role, although they remain involved in more than one contributorship type.

When comparing archetypes by number of publications (Figure 5B), we observe almost no differences on publication rates in the junior stage. Nonetheless, differences emerge for later career stages. Except for the late-career stage, where Supporting scientists are the most productive, the Leader archetype exhibits higher productivity, followed by Supporting. Specialized scientists appear to be much less productive than scientists assigned to the other two archetypes in the early- and mid-career stages. This pattern is also observed for Specialized, in the case of citation impact. However, differences in terms of share of highly cited publications between the Leader and the Supporting archetypes are much smaller, with the latter exhibiting higher values.

Archetypes and gender. Figure 6 shows that scientists are unevenly distributed by gender in each archetype. Note that scientists from different generations are included in the analysis, therefore, caution should be expressed in drawing any conclusion on the number of scientists by gender that reach the late-career stage. The share of women who reach the late-career stage is affected by the generational diversity of sci-

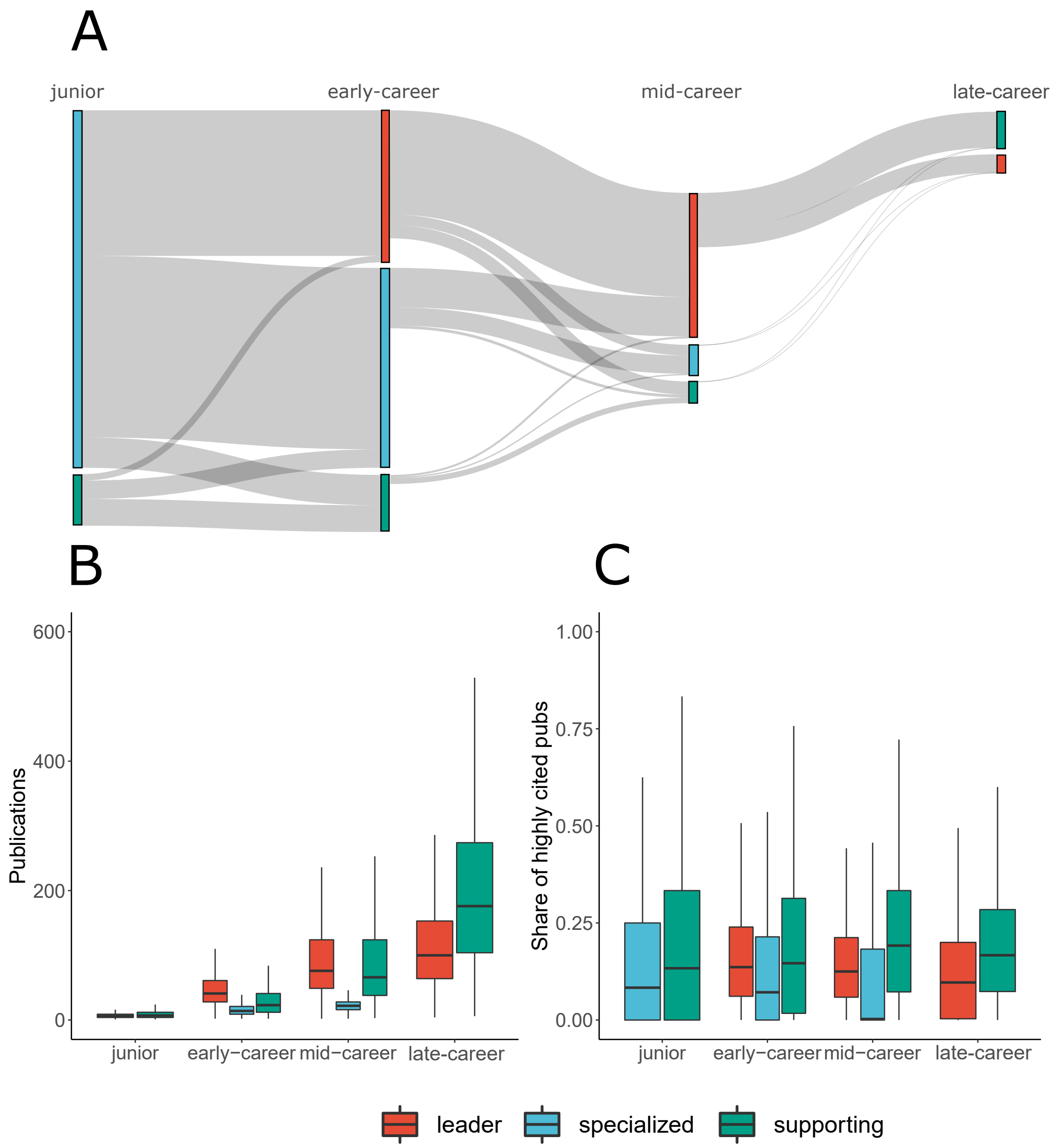


Fig. 5. Career trajectories Sankey diagrams **(A)**, productivity **(B)**, and citation impact **(C)** boxplots by archetype and career stage. A Sankey diagram of scientists assigned to each archetype at each career stage, indicating shifts between stage. **Blue** refers to the Specialized profile, **Red** refers to the Leader profile and **Green** refers to the Supporting profile.

entists and hence we make comparisons only within career stage. As observed, a gender disparity on the distribution by archetype and stage is consistent in all career stages. The share of men is higher for the Specialized archetype at the junior stage, and for the Leader archetype at the early- and mid-career stages. The second most frequent is the Specialized archetype, with few men in the Supporting archetype, except for the late-career stage. Women are less likely to appear as the Leader archetype in the early- and mid-career stages. Whereas 87% of men in the junior stage have a Specialized archetype, 43% and 77% in the early- and mid-stage are designated as Leaders; 84% of women in junior are Specialized, and only 27% and 65% in early- and mid-career stages are assigned to that profile. The gender distribution becomes more balanced again at the late-career stage, where 35% of men and 31% of women are in the Leader archetype. In summary, women appear to group within the Specialized archetype in the early-career stage, and show similar distributions to that of men at the other career stages, although the shares of the Leader archetype are consistently lower to that of men.

Archetypes and author position. We analyze the relationship between author order and archetypes by career stage. Figure 7 shows the share of papers by archetype and career stage of scientists based on their author position. Middle authorships occupy a larger share of publications irrespective of the archetype or career stage, which is a consequence of the fact that any paper with more than three authors, most authors are in middle positions. We do observe, however, variation in middle authorship by career stage. At the junior stage, middle authorships account for half of the papers from Specialized scientists, while Supporting scientists occupy a middle position in almost 75% of their publications. In the early-career stage, the Leader archetype emerges, exhibiting

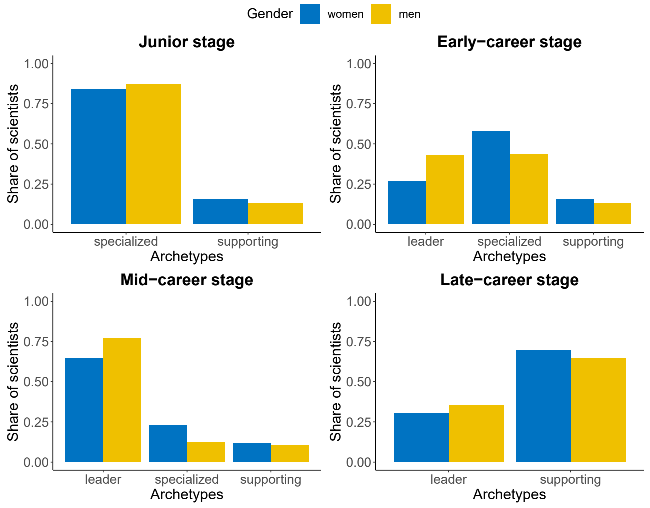


Fig. 6. Percentage of scientists by gender and career stage for each archetype.

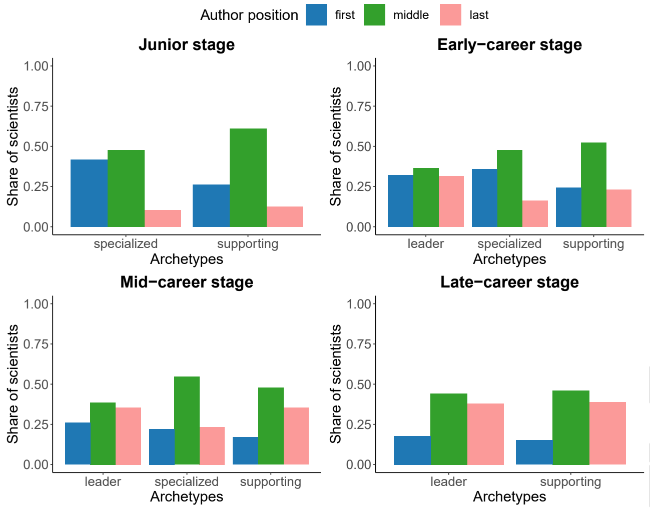


Fig. 7. Share of publications by author position for each archetype and career stage.

a more balanced share of publications between first (32%), middle (37%) and last positions (32%). Specialized scientists publish a slightly higher share as first authors (36%) but almost in half of their papers appear in middle positions (48%). The Supporting archetype publishes more than half of their papers as middle authors (53%), evenly distributed between first and last authored publications.

At the mid-career stage, Leader scientists start to shift to last positions (36%), with only 26% of their publications being first authored. Specialized scientists become the middle authors in 55% of their publications and are last authors on 23% of their publications. Supporting scientists, however, position themselves as last authors in 35% of their publications. The Specialized archetype disappears in the late-career stage. The Leader and Supporting archetypes show similar distributions of publications according to their author position, revealing that at this stage, author position is more related with seniority than contributorship.

Discussion

Scientists are immersed in a reward system that evaluates them individually following uniform expectations of leadership and excellence (1, 15, 16, 31). Such evaluation mechanisms do not consider the fact that science increasingly relies on larger teams of scientists to be able to tackle grand challenges (14), which requires specialization across tasks (12). Claims have been made that the breadth and diversity of profiles need to be taken into account when assessing individual research performance (8, 10). Here we identify and characterize such diversity of profiles by career stage, by combining contribution statements with bibliometric variables and applying a machine learning algorithm to predict contributions.

We find that scientists display different archetypes at different stages, following many paths during their career trajectory. Some paths, however, come at a cost. Out of the 222,295 scientists included in our dataset, only 12% reached the late-career stage, out of which the vast majority (98%) displayed a Specialized archetype in their junior stage. Even though most of them belonged to the Leader archetype in their early- and mid-career stages, scientists at the late-career stage mostly exhibit a Supporting archetype (66%). This could be happening because many scientists adopt a secondary role when they reach seniority, leaving the leading role to their younger colleagues.

The names assigned to each archetype are figurative but reflect an implicit hierarchy in science. This hierarchy exists at each career stage, indicating that the diversity of profiles is not the result of scientists evolving in their career trajectory and adopting different roles, but that diverse archetypes exist between and within career stages. The archetypal analysis identified no Leader archetype at the junior stage, when scientists are still 'earning their stripes', nor are there Specialized scientists in the late-career stage. Such reality enters into conflict with the current expectations on research careers, which consider roles to be attached with career stages and steps that must be made to progress. Our findings have important policy implications as they indicate that scientists' career design may be at odds with the way science is produced, and suggest a complete reform wherein reproduction of Leaders is not the only model of success (8).

Our results demonstrate the high versatility of the Leader archetype: scientists with this profile are able to move seamlessly across archetypes during their careers. While there are some scientists with a Specialized or Supporting profile who manage to shift to the other three archetypes, most of the scientists fitting these archetypes in our dataset do not progress to more senior stages. Our analysis on productivity and citation impact by archetype sheds light on the mechanisms which may be affecting trajectories. Specialized scientists are less productive and have a lower share of highly cited publications than Leaders and Supporting scientists, which may serve a disadvantage for career advancement in environments which prioritize bibliometric indicators in research assessment (Figure 5B,C). The lack of assessment schemes sensitive to the diversity of profiles, partly due the inappropriate use of bibliometric indicators at the individual researcher level (1, 32), limits the capacity of policies to correct for inequalities observed across and within archetypes. Structural changes in the academic reward system are necessary to support the advancement and retention of Specialized and Supporting scientists.

We observe consistent differences in the distribution of archetypes by gender, which may contribute to explain the higher rates of attrition for women (21). Early-career stage is key to the development of scientific careers, and it is at this stage that large gender differences are observed. While in the other career stages women and men exhibit a similar distribution of archetypes, women are more likely to be of the Specialized archetype in early-career, while men are more likely to be Leaders. That women disproportionately engage in technical labor–even when controlling for academic age–has been demonstrated in previous studies (18). This is consistent with general patterns in academic labor; for example, the higher service work done by women academics (33).

Contributorships are generally associated with author order (10, 25), based on the presumption that first and last author will have major roles, while middle authors will play a secondary role. These roles reinforce hierarchy and organizational strategies: leaders set the agenda and define lines of work, whereas technicians are prized for their ability to implement this agenda (34). This model, however, does not provide equal access to career advancement for all scientists: those showcasing a Specialized or Supporting archetype in their early- and mid-career stages have greater difficulties to progress in their research career. These obstacles affect women at a greater extent than men, as a higher proportion of female scientists adopt these roles. Our findings suggest systematic biases on the selection of individuals which may be hampering the efficiency of the scientific system to self-organize itself and assemble robust and diverse scientific teams.

Methods and Materials

The data needed to reproduce the paper are openly accessible at http://doi.org/10.5281/zenodo.3891055.

Our analysis is based on two datasets: a seed dataset of contributorship statements and dataset of researchers' late-publication histories. The seed dataset combines bibliometric and contributorship data for 85,260 publications from 7 PLOS journals, during the 2006-2013 period. Although many biomedical journals have adopted contributorship statements (e.g., BMJ, The Lancet), PLOS journals provide data in an XML format which ease the data retrieval process.

This dataset is used to train a predicting model of contributorship based on bibliometric variables. The full publication histories dataset contains the complete publication history of the 222,925 authors selected from the list of publications of the first dataset. This dataset is used to predict authors' contributorship per paper and is later aggregated at the individual level to identify archetypes of scientists per career stage.

Contributorship statements. We used a dataset of 85,260 distinct PLOS papers published during the 2006-2013 period. This dataset was gathered from the PLOS website in combination with Web of Science data. Full account of the complete extraction procedure is provided in a previous study (12). For each publication and author, a dummy value is assigned based on the tasks they performed. Table 2 shows the list of journals together with the number of publications per journal. 88% of the publications have been published in PLOS One. Seven types of contributions were originally included in the dataset. Only five of those contributorships are being used consistently throughout the dataset. "Approved final version of the manuscript" and "Other contributions" are present in less than 5% and 20% of the papers respectively. While the former is a requirement of the ICMJE and therefore is used mostly in PLOS Medicine, the latter is not an individual category, but an aggregate containing nearly 20,000 different types of contributions. The low incidence of the "Approved final version" contribution together with the difficulties in interpreting the "Other" contributorship led to their exclusion from the analysis.

Bibliometric data. The bibliometric data is obtained from the CWTS (Leiden University) in-house version of the Web of Science. This database contained at the moment of analysis all publications included in the Science Citation Index Expanded, Social Science Citation Index, and Arts and Humanities Citation Index for the 1980-2017 period. Furthermore, an author name disambiguation algorithm (23) is applied to the complete database, allowing to identify a scientist's complete publication history. This allowed us to retrieve, for each paper contained in the contribution dataset, bibliometric variables at the publication and at the author level. A set of seven bibliometric variables is considered, which is described in Table 1 by author-publication combination. Here, we highlight the use of the variable years since first publication (YE). This variable is used to determine the age of scientists and is used later to estimate the different career stages of the individuals identified. Our use of the year of first publication as an indicator for academic age is based on previous research (35), in which the year of first publication is found to be the best predictor for the age of scientists. In the case of productivity, we use a full counting approach. While fractional counting can be considered as being more accurate from a mathematical point of view (36), the focus here is on the previous publication experience of the author and how that might influence their role in future publications. Hence we consider full counting to suit best the purposes of the analysis.

Merging of bibliometric and contribution data. The merging process was undertaken by matching documents by their DOI identifier and authors who had the same initials and surname in both datasets. We only included papers for which all authors were successfully matched. After this process was undertaken, we ended up with a total of 77,749 publications, containing a total of 369,537 disambiguated unique authors.

Subject field identification. We assigned a subject field to each publication and filtered only those publications that belong to the Medical and Life Sciences to ensure consistency

Table 1. Definition of variables included in the dataset.

Acronym	Definition	Source
Bibliometric variables		
PO	Author's position in the paper	WoS
AU	Total number of authors in the paper	WoS
DT	Document type. Letters are excluded	WoS
CO	Number of countries to which authors of the paper are affiliated	WoS
IN	Number of institutions to which authors of the paper are affiliated	WoS
YE	Number of years since first publication at the time the paper was published	WoS
PU	Average number of publications (full counting) per year of the author at the time the paper was published	WoS
Contribution variables		
WR	Wrote the paper	PLOS
AD	Analyzed the data	PLOS
PE	Performed the experiments	PLOS
CE	Conceived and designed the experiments	PLOS
CT	Contributed reagents/materials/analysis tools	PLOS
NC	Number of contributions	PLOS

Table 2. Distribution of papers by journal of the seed dataset on contributions.

Journal	No. of papers
PLOS ONE	62,174
PLOS GENETICS	2,408
PLOS PATHOGENS	1,882
PLOS COMPUTATIONAL BIOLOGY	1,684
PLOS NEGLECTED TROPICAL DISEASES	1,432
PLOS BIOLOGY	697
PLOS MEDICINE	417

on publication patterns and distribution of contributorships. For this, we used the Dutch NOWT Classification which introduces three levels of categorization: 7 broad areas, 14 fields, and 34 subjects. This classification is linked to the the Web of Science subject categories (see correspondence here https://www.cwts.nl/pdf/nowt$_{c}lassification_{s}c.pdf$). The classification is made at the journal level, which implies that, given the high incidence of the PloS One papers in our data set, most publications would be categorized as Multidisciplinary. To overcome this issue, publications in Multidisciplinary were reclassified into other more specific fields based on their reference lists. We identified the journal to which each of the references of the publications in our data set belong to. Then, we assigned to each publication the field from which most of its references come from. Finally, we only include those which are assigned to the Medical and Life Sciences fields. A total of 70,694 publications and 347,136 distinct authors were extracted from this process, constituting the "seed data set".

Publication history of individual scientists. We reconstructed the publication histories of scientists, and predicted their contributions throughout their careers. The set of authors identified is retrieved from the seed dataset to ensure consistency on the predictions of the Bayesian Network model. But a series of thresholds are imposed. First, we retrieve authors' gender using the following sources to identify gender: Gender API, Genderize.io and Gender Guesser. We apply a 90% accuracy threshold before assigning gender and only include those authors who surpass such threshold. Second, we include only authors whose first publication occurred from 1980 onwards. While the CWTS in-house database includes publications prior to 1980, it does not contain metadata of sufficient quality as to rely on the name disambiguation algorithm. Hence, authors with their first publication prior to 1980 are discarded. Third, we include only authors who have contributed to at least five publications. We do this for two reasons. On the one hand, we remove transient authors, that is, those who have published sporadically, and focus only on scientists that have more chances of being pursuing a research career. On the other hand, this increases the accuracy of the author name disambiguation performed on those researchers. This is specially relevant since the algorithm adopts a conservative approach: when confronted with individuals having outlier patterns of behavior, such as rapid shifts across publication venues, disciplines and co-authors, it will consider them as different authors and consequently split their publications across different "individuals". Hence, by including a publication threshold, we focus on those individuals for whom the algorithm is more robust and accurate at identifying them uniquely. Last, we remove the publications classified as letters to ensure consistency between the two datasets with respect to the document type. As a result, the final dataset contains a total of 222,925 individuals and 6,236,239 distinct publications. The reason for the much larger set of publications is that for those scientists identified in the Seed dataset, we have expanded to all their other publications identified by the algorithm (and not just those from Table 2).

Bayesian networks for predicting contributorships. Bayesian networks (BNs) graphically depict interactions among dependent multivariate data. The network structure represents a directed acyclic graph (DAG), where nodes represent random variables and arcs encode direct influences. Along with dependence statements, a BN encodes conditional independence statements among random variables. These conditional independencies are described by the d-separation concept (28) and are captured graphically by the BN structure. The Markov property ensures a convenient factorization of the joint distribution of the multivariate data. Say n continuously distributed random variables $X_1, X_2, \ldots, X_n$ are modeled by a Bayesian network. Then, the joint probability density function can factorize in the following manner

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i | P_a(X_i)), \qquad (1)$$

where $Pa(X_i)$, for $i = 1, \ldots, n$, represents the parent set of node X_i, that is, the set of nodes (variables) whose arcs are directed at X_i. The conditional densities $f(x_i|P_a(X_i))$, for $i = 1, \ldots, n$, of each random variable conditioned on its set of parent nodes encode the Markov property. The joint density factorization therefore depends on the structure of the network, that is, on the presence or absence of arcs and their directions.

There are numerous structures that can be considered, and the number of structures grows super-exponentially with the number of variables (35). Let a_n denote the number of BNs with n random variables. Then

$$a_n = \sum_{k=1}^{n} (-1)^{k+1} (\frac{n}{k}) 2^{k(n-k)} a_{n-k}, \qquad (2)$$

where $a_0 = 1$. The structure of a BN can be learned from data or from experts, or from mixing data-driven algorithms with expert input.

Data driven learning algorithms of a BN structured are broadly categorized into constraint-based and score-based learning algorithms (36). Constraint-based methods rely on conditional independence tests, whereas score-based methods employ likelihood-based metrics to evaluate structures. Both types of algorithms also contain a search procedure, such as a local search in the space of network structure (36, 37). We employ the Max-Min Hill-Climbing (MMHC) algorithm (29), which combined techniques from constraint and score-based algorithms, along with an initial local discovery algorithm of edges without any orientation.

We have employed a mixed approach, which imposed, via a white list, the direct influences of bibliometric to contributorship variables. The white-listed arcs are depicted in red in Figure S1. It is noteworthy that the arcs were present from employing the MMHC data-driven algorithm, and only the direction was switched. These white-listed nodes have been accounted for in learning the structure with the remaining variables. Thus, the remaining arcs in the BN together with their directionality have been fully assigned by using data-driven algorithms.

Finally, the BN structure has been subjected a robustness check by employing a bootstrap procedure, by which bootstrap replications of the data have been sampled 50 times from the initial data, with replacement. The bootstrap samples had the same size as the initial dataset. The MMHC algorithm has provided network structures and the arcs that have appeared in at least 80% of the structures have been retained. Figure 2B illustrates the resulting network. The BN analysis has been performed using the bnlearn package in R (36).

Table 3. Classification error rates from cross-validation of Bayesian Network model for the contribution variables.

Variables	Min.	Median	Mean	Max.
WR	0.062	0.064	0.064	0.065
AD	0.064	0.067	0.067	0.069
PE	0.072	0.075	0.075	0.077
CE	0.062	0.064	0.064	0.066
CT	0.077	0.078	0.078	0.081
NC	0.729	0.732	0.732	0.735

Crossvalidation. To validate the BN used for predictions, we perform a k-fold cross-validation. The data are split in 10 subsets. For each subset, in turn, the BN is fitted on the other k - 1 subsets and a predictive loss function is then computed using that subset. Loss estimates for each of the k subsets are then combined to give an overall predictive loss. Since we are interested in predicting whether a scientist had a certain contributorship for the publications in the dataset, we translate the predictive loss into classification error. That is, we quantify the classification error rate of the BN in predicting a certain contributorship, given the bibliometric information of scientists and publications. The classification error rates obtained for each contributorship are shown in Table 3. While the error rates obtained are quite low, it is true that this validation is performed using data which is of the same nature as the data on which the BN has been quantified. This means that the extent to which contribution patterns in our dataset can be inferred to other datasets should be further investigated using different journals or fields.

Constructing scientific profiles.

Data aggregation. Predicted probabilities of all contributorship types obtained from the BN are available for each author-publication combination. We aim to aggregate those prediction at the author level, that is, to derive, for each scientist, the probability of fulfilling each contribution role. For this, we used the median probability value per contribution type. Furthermore, we grouped the publications by career stage, that is, publications within 5 years from the first publication (junior stage), publications between 5 and 15 years from first publication (early-career), publications between 15 and 30 years from first publication (mid-career) and publications after 30 years from first publication (late-career). Here must note that the selection of the time periods was selected for convenience and that any other division could have been selected. For each researcher, we obtain a median probability per contribution type and career stage.

Suppose within career stage i, with $i = 1, \ldots, 4$, a scientist has k publications. Let p^i_j the probability that the scientist performs contributorship j within career stage i, for $j = 1, \ldots, 5$

denoting the five different types of contributions (WR, CT, CE, PE, AD). Then

$$p_j^i = Median(p_{j,1}^i, p_{j,2}^i, \dots, p_{j,k}^i), \quad \textbf{(3)}$$

where $p_(j,1)^i$ is the predicted probability of contribution j of the scientist's first publication in career stage i. For the number of contributions (NC), the same aggregation is applied

$$NC_j^i = Median(NC_1^i, \dots, NC_k^i) \quad \textbf{(4)}$$

where NC_1^i is the predicted number of contributions for the first paper in career stage i.

Robust archetypal analysis. Profiles of researchers, by career stage, are obtained using a robust archetypal analysis. Archetypal analysis aims to identify archetypes that emerge from the given contribution data for scientists. This approach has been previously applied to identify scientists' profiles based on citation and publication data (40). The archetypes are extreme observations in a multivariate dataset and represented as convex combinations of the observations in the dataset that result from a least squares problem (24). For multivariate data with n observations (scientists, per career stage, in our case) and m random variables (types of contributorships, in our case), then X is a n×m matrix denoting the aggregated dataset. For given k archetypes, denote by Z the k×m the matrix of archetypes, represented in terms of the types of contributorships. Then, the residual sum of squares (RSS) plotted in Figure S1 is denoted by

$$RSS = ||X - aZ^T||_2, \quad \textbf{(5)}$$

with $Z = X^T\beta$, where α,β are positive coefficients and where $||\cdot||_2$ denotes the Euclidean matrix norm. In turn, each observation in the dataset can be represented as a convex combination of the archetypes

$$X \approx aZ^T \quad \textbf{(6)}$$

In the standard approach of archetypal analysis, each residual contributes to the RSS with equal weight. The archetypal analysis is thus sensitive to outliers, whose large residuals can contribute significantly to the RSS. A robust archetypal analysis (30) has been proposed to weight down the influence of outliers to the construction of archetypes. By letting W be a n×n matrix of weights, we define the weighted RSS

$$RSS = || W(X - \alpha Z^T) ||_2 \,. \quad \textbf{(7)}$$

The weights can be chosen by the user or can be chosen to depend on each observation's residual. The robust archetypal analysis proposed by (30) proposes an iterative re-weighted least squares algorithm. Unlike the k-means clustering approach, which engages averaging when profiling out clusters, archetypal analysis focuses on extremes and explore the heterogeneity of complex multivariate data. Furthermore, archetypes are not forced to be mutually exclusive, as principal components are, nor do they remain the same when the number of considered archetypes is changing. The archetypal analysis has been performed using the archetypes package in R (38).

Limitations of the study.

Representativeness of the sample of scientists. The analysis is based on a set of publications and a sample of scientists which may not represent accurately the whole population of scientists. This means that, despite the robustness of the results, any inference to the whole population should be done with caution. Furthermore, the thresholds imposed to introduce such scientists in the archetypal analysis further restricts such inference endeavour. If we compare the productivity distributions of our set of researchers and for the whole population of the Web of Science, we observe that while we still retain a high skewness of productivity, this is much lower than the overall one.

Identification of scientists. The study relies heavily on the competence of an author name disambiguation algorithm to correctly identify disambiguated authors. As previously noted, this algorithm has some limitations which are partially overcome by the production thresholds imposed. However, inaccurate assignments can still occur.

Author age. We estimate researchers' age based on the year of first publication and build the four career stages based on such year. However, alternative approaches could have been adopted and these could have some impact on the results. For instance, first year of first-authored publication could have been used instead. The selection of the first year of publication is based on empirical data suggesting that it is the best proxy for PhD year (39).

Taxonomy of contributorships. In this paper, contributions are classified into five types. These types are obtained from the data itself. However, one may question the appropriateness of the number and contribution types. The ones used in this paper are consistent with those used in other studies (12), but different from those proposed in the recent CRediT initiative, which defines up to 14 types of contributions. Furthermore, evidence suggests that author self-reporting on contributorship is not exempt of limitations. Questions like the extent to which contribution types are field-dependent are still unsolved. With this respect, our predictions already point towards some of these issues. Despite the low error rates, we observe that the distribution of predicted probabilities exhibits a normal distribution for writing the manuscript (Figure 2A). This could be due to the ambiguity of the statement. As observed in the CRediT intitiative, this statement is disclosed into two: wrote the first draft and wrote parts of the manuscript and revised. Such distinction might help the model to better discriminate contributorships.

ACKNOWLEDGEMENTS

The authors thank Lidia Carballo-Costa and Alejandro Quintela del Río (Universidade da Coruña, Spain) for their support in applying the archetypal analysis. This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 707404 and the South African DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP). The authors declare that they have no competing interests. Data are derived from Clarivate Analytics Web of Science. © Copyright Clarivate Analytics 2020. All rights reserved.

Bibliography

1. Erin C. McKiernan, Lesley A. Schimanski, Carol Muñoz Nieves, Lisa Matthias, Meredith T. Niles, and Juan Pablo Alperin. Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations. *PeerJ Preprints*, (e27638v2), April 2019. doi: 10.7287/peerj.preprints.27638v2.
2. David Moher, Florian Naudet, Ioana A. Cristea, Frank Miedema, John P. A. Ioannidis, and Steven N. Goodman. Assessing scientists for hiring, promotion, and tenure. *PLOS Biology*, 16(3):e2004089, March 2018. ISSN 1545-7885. doi: 10.1371/journal.pbio.2004089.
3. Samuel F. Way, Allison C. Morgan, Daniel B. Larremore, and Aaron Clauset. Productivity, prominence, and the effects of academic environment. *Proceedings of the National Academy of Sciences*, 116(22):10729–10733, May 2019. ISSN 0027-8424, 1091-6490. doi: 10/ggbv2b.
4. Hua-Wei Shen and Albert-László Barabási. Collective credit allocation in science. *Proceedings of the National Academy of Sciences*, 111(34):12325–12330, 2014.
5. Barbara F. Reskin. Academic Sponsorship and Scientists' Careers. *Sociology of Education*, 52(3):129–146, 1979. ISSN 0038-0407. doi: 10/cx47c5.
6. Filippo Radicchi, Santo Fortunato, Benjamin Markines, and Alessandro Vespignani. Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, 80(5):056103, November 2009. doi: 10/b7k4hd.
7. Alexander Michael Petersen, Santo Fortunato, Raj K. Pan, Kimmo Kaski, Orion Penner, Armando Rungi, Massimo Riccaboni, H. Eugene Stanley, and Fabio Pammolli. Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences*, 111 (43):15316–15321, 2014. doi: 10.1073/pnas.1323111111.
8. Staša Milojević, Filippo Radicchi, and John P. Walsh. Changing demographics of scientific careers: The rise of the temporary workforce. *Proceedings of the National Academy of Sciences*, 115(50):12616–12623, December 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1800478115.
9. Roger Guimerà, Brian Uzzi, Jarrett Spiro, and Luís A. Nunes Amaral. Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, 308 (5722):697–702, April 2005. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1106340.
10. Philippe Mongeon, Elise Smith, Bruno Joyal, and Vincent Larivière. The rise of the middle author: Investigating collaboration and division of labor in biomedical research using partial alphabetical authorship. *PLOS ONE*, 12(9):e0184601, September 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0184601.
11. Staša Milojević. Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences*, 111(11):3984–3989, March 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1309723111.
12. Vincent Larivière, Nadine Desrochers, Benoît Macaluso, Philippe Mongeon, Adèle Paul-Hus, and Cassidy R Sugimoto. Contributorship and division of labor in knowledge production. *Social Studies of Science*, 46(3):417–435, June 2016. ISSN 0306-3127. doi: 10.1177/0306312716650046.
13. M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5200–5205, June 2004. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0307545100.
14. Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. The Increasing Dominance of Teams in Production of Knowledge. *Science*, 316(5827):1036–1039, May 2007. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1136099.
15. Thijs Bol, Mathijs de Vaan, and Arnout van de Rijt. The Matthew effect in science funding. *Proceedings of the National Academy of Sciences*, page 201719557, April 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1719557115.
16. Robert K. Merton. The Matthew Effect in Science The reward and communication systems of science are considered. *Science*, 159(3810):56–63, May 1968. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.159.3810.56.
17. Jonathan R. Cole and Harriet Zuckerman. The productivity puzzle: persistence and change in patterns of publication of men and women scientists,. Advances in Motiva-tion and Achievement. *A Research Journal Women in Science*, 2:217–258, 1984.
18. Benoit Macaluso, Vincent Larivière, Thomas Sugimoto, and Cassidy R. Sugimoto. Is Science Built on the Shoulders of Women? A Study of Gender Differences in Contributorship. *Academic Medicine*, 91(8):1136–1142, August 2016. doi: 10.1097/ACM.0000000000001261.
19. Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R. Sugimoto. Bibliometrics: Global gender disparities in science. *Nature*, 504(7479):211–213, December 2013. ISSN 1476-4687. doi: 10/qgf.
20. Patrick Gaule and Mario Piacentini. An advisor like me? Advisor gender and post-graduate careers in science. *Research Policy*, 47(4):805–813, May 2018. ISSN 0048-7333. doi: 10.1016/j.respol.2018.02.011.
21. Junming Huang, Alexander J. Gates, Roberta Sinatra, and Albert-Laszlo Barabasi. Historical comparison of gender inequality in scientific careers across countries and disciplines. *arXiv:1907.04103 [physics]*, July 2019. arXiv: 1907.04103.
22. Grit Laudel and Jochen Gläser. From apprentice to colleague: The metamorphosis of Early Career Researchers. *Higher Education*, 55(3):387–406, March 2008. ISSN 1573-174X. doi: 10.1007/s10734-007-9063-7. Citation Key Alias: laudelApprenticeColleagueMetamorphosis2007.
23. Emiel Caron and Nees Jan van Eck. Large scale author name disambiguation using rule-based scoring and clustering. In *19th International Conference on Science and Technology Indicators."Context counts: Pathways to master big data and little data"*, pages 79–86. CWTS-Leiden University Leiden, 2014.
24. Adele Cutler and Leo Breiman. Archetypal Analysis. *Technometrics*, 36(4):338–347, November 1994. ISSN 0040-1706. doi: 10.1080/00401706.1994.10485840.
25. Henry Sauermann and Carolin Haeussler. Authorship and contribution disclosures. *Science Advances*, 3(11):e1700404, November 2017. ISSN 2375-2548. doi: 10.1126/sciadv.1700404. tex.ids: sauermannAuthorshipContributionDisclosures2017a.
26. Zaida Chinchilla-Rodríguez, Cassidy R. Sugimoto, and Vincent Larivière. Follow the leader: On the relationship between leadership and scholarly impact in international collaborations. *PLOS ONE*, 14(6):e0218309, June 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0218309.
27. International Committee of Medical Journal Editors. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals (icmje recommendations). Technical report, International Committee of Medical Journal Editors, 2015.
28. Thomas Dyhre Nielsen and Finn Verner Jensen. *Bayesian networks and decision graphs*. Springer Science & Business Media, 2009.
29. Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, October 2006. ISSN 1573-0565. doi: 10.1007/s10994-006-6889-7.
30. Manuel J. A. Eugster and Friedrich Leisch. Weighted and robust archetypal analysis. *Computational Statistics & Data Analysis*, 55(3):1215–1225, March 2011. ISSN 0167-9473. doi: 10/fsbb9k.
31. Barbara F. Reskin. Scientific Productivity and the Reward Structure of Science. *American Sociological Review*, 42(3):491–504, 1977. ISSN 0003-1224. doi: 10.2307/2094753.
32. Diana Hicks, Paul Wouters, Ludo Waltman, Sarah De Rijcke, and Ismael Rafols. Bibliometrics: the leiden manifesto for research metrics. *Nature*, 520(7548):429–431, 2015.
33. Thamar M Heijstra, Þorgerður Einarsdóttir, Gyða M Pétursdóttir, and Finnborg S Steinþórsdóttir. Testing the concept of academic housework in a european setting: Part of academic career-making or gendered barrier to the top? *European Educational Research Journal*, 16 (2-3):200–214, 2017.
34. Bruno Latour and Steve Woolgar. *Laboratory life: The construction of scientific facts*. Princeton University Press, 2013.
35. Robert W Robinson. Combinatorial mathematics. In C. H. C. Little, editor, *Lecture Notes in Mathematics*, pages 28–43. Springer-Verlag, 1977.
36. Marco Scutari and Jean-Baptiste Denis. *Bayesian Networks: With Examples in R*. Chapman and Hall/CRC, Boca Raton, Fla., 1 edition edition, July 2014. ISBN 978-1-4822-2558-7.
37. Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
38. Manuel J. A. Eugster and Friedrich Leisch. From Spider-Man to Hero &mdash; Archetypal Analysis in R. *Journal of Statistical Software*, 30(1):1–23, April 2009. ISSN 1548-7660. doi: 10/ggck2c.
39. Gabriela F. Nane, Vincent Larivière, and Rodrigo Costas. Predicting the age of researchers using bibliometric data. *Journal of Informetrics*, 11(3):713–729, August 2017. ISSN 1751-1577. doi: 10.1016/j.joi.2017.05.002.