PNAS
www.pnas.org

Supplementary Information for

Gender Inequities in the Online Dissemination of Scholars' Work

Orsolya Vásárhelyi, Igor Zakhlebin, Staša Milojević, and Emőke-Ágnes Horvát

Emőke-Ágnes Horvát
E-mail: a-horvat@northwestern.edu

This PDF file includes:

References

1. R Oaxaca, Male-female wage differentials in urban labor markets. *Int. economic review*, 693–709 (1973).

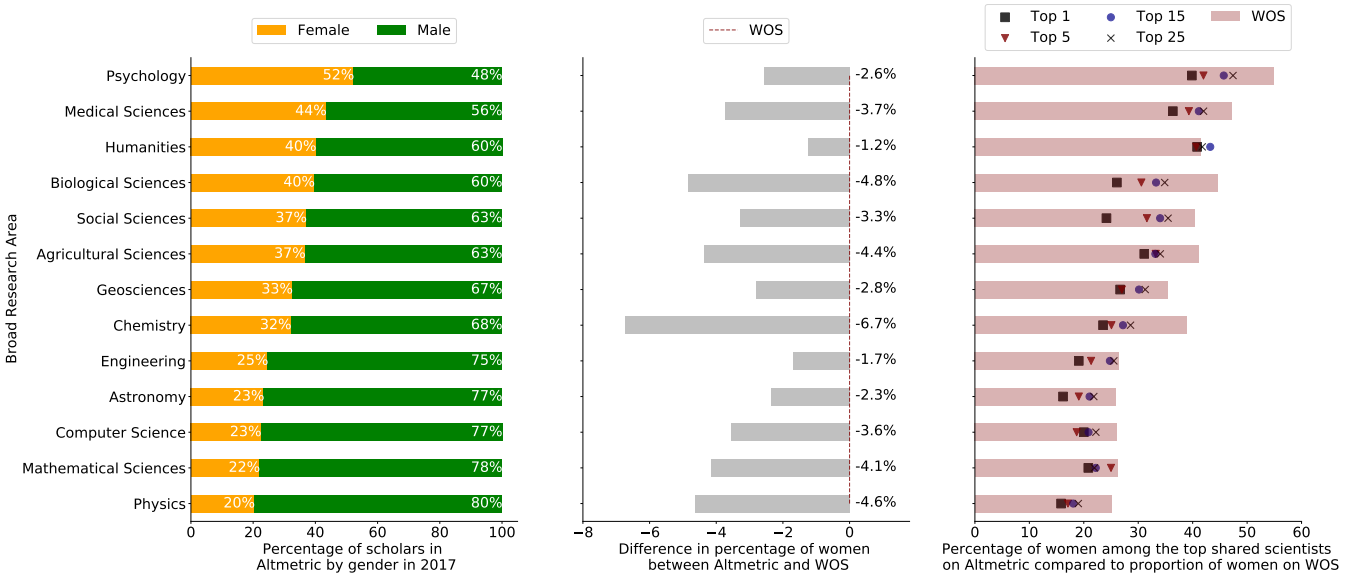

Fig. S1. Online success of female scholars in various broad research areas in 2017. The figure is based on 1,034,476 articles written by 4,086,476 scientists. Left: Percentage of women among scholars who had articles mentioned online in 2017. Middle: Online representation of female scholars based on Altmetric in comparison with the ratio of women who published research papers in 2017 according to the Web of Science (WoS). Right: Proportion of women in the top 1, 5, 15, and 25% of the scientists with the most mentions online compared with percentage of women who published according to the WoS.

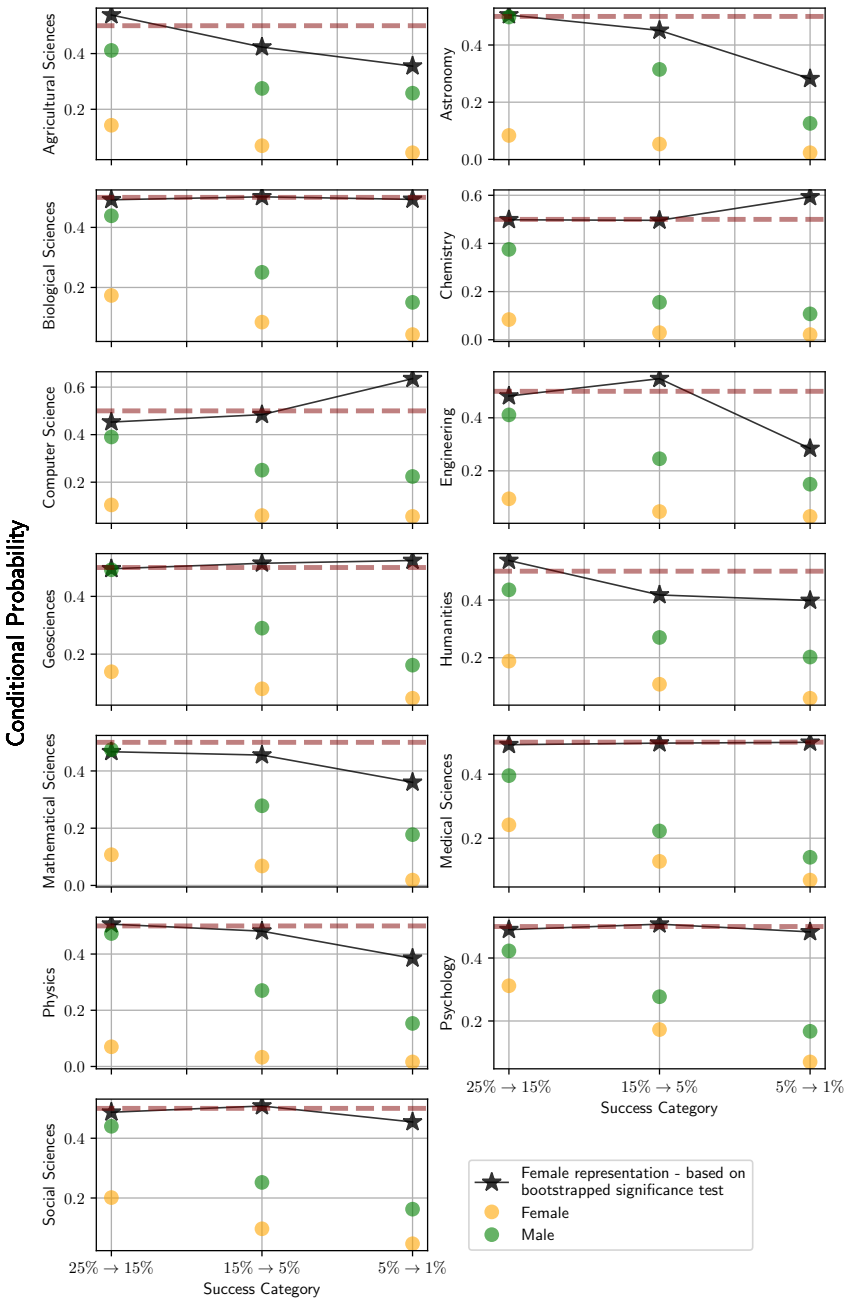

Fig. S2. Results of the bootstrapped significance test that evaluates the conditional probability that female and male scientists are in a higher online success category (top 15%, 5%, 1%) given their presence in a lower category (top 25%, 15%, 5%) in the studied broad scientific areas.

Broad Research Area	Intercept	Scientific Impact	Female: Scientific Impact	Social Capital	Female: Social Capital	Network Femaleness	Female: Network Femaleness	Network Maleness	Female: Network Maleness	R^2	N
Agricultural Sciences	0.29	1.11	1.56	1.88	0.57	1.46	0.85	1.90	0.58	0.22	1,458
	(0.00)	(0.06)	(0.00)	(0.00)	(0.00)	(0.00)	(0.11)	(0.00)	(0.00)		
Astronomy	0.21	1.55	1.18	0.90	0.77	1.15	0.96	1.43	0.77	0.16	2678
	(0.00)	(0.00)	(0.07)	(0.02)	(0.02)	(0.01)	(0.54)	(0.00)	(0.01)		
Biological Sciences	0.50	1.63	1.31	1.24	0.90	1.44	0.85	1.59	0.79	0.21	61,495
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)		
Chemistry	0.11	1.41	1.24	1.40	0.99	1.11	1.04	1.47	0.68	0.31	17438
	(0.00)	(0.00)	(0.00)	(0.00)	(0.82)	(0.00)	(0.26)	(0.00)	(0.00)		
Computer Science	0.28	1.01	0.77	1.45	1.36	1.50	0.96	1.63	0.97	0.19	1,510
	(0.00)	(0.78)	(0.05)	(0.00)	(0.04)	(0.00)	(0.70)	(0.00)	(0.80		
Engineering	0.14	1.42	1.13	1.44	1.15	1.36	0.84	1.55	0.68	0.22	2,962
	(0.00)	(0.00)	(0.24	(0.00)	(0.22)	(0.00)	(0.01)	(0.00)	(0.00)		
Geosciences	0.42	1.71	1.39	0.99	0.95	1.29	0.87	1.38	0.67	0.17	8,790
	(0.00)	(0.00)	(0.00)	(0.79)	(0.30)	(0.00)	(0.00)	(0.00)	(0.00)		
Mathematical Sciences	0.26	1.03	1.14	2.34	1.04	1.34	0.97	1.58	0.70	0.23	1,062
	(0.00)	(0.62)	(0.58)	(0.00)	(0.87)	(0.00)	(0.81)	(0.00)	(0.06)		
Medical Sciences	0.41	1.24	1.11	1.20	0.99	1.30	0.97	1.24	0.86	0.19	121,462
	(0.00)	(0.00)	(0.00)	(0.00)	(0.29)	(0.00)	(0.01)	(0.00)	(0.00)		
Physics	0.33	1.63	1.07	1.30	1.01	1.14	1.05	1.64	0.69	0.22	9,287
	(0.00)	(0.00)	(0.29)	(0.00)	(0.90)	(0.00)	(0.22)	(0.00)	(0.00)		
Psychology	1.18	1.24	1.12	1.60	0.96	1.21	1.01	1.29	0.86	0.18	10,670
	(0.00)	(0.00)	(0.02)	(0.00)	(0.47)	(0.00)	(0.81)	(0.00)	(0.00)		
Social Sciences	0.56	1.25	1.08	1.74	1.03	1.13	1.05	1.16	0.81	0.15	6,424
	(0.00)	(0.00)	(0.16)	(0.00)	(0.69)	(0.00)	(0.21)	(0.00)	(0.00)		

Table S1. Odds ratio and significance level of variables in logistic regression models to predict presence among the top 25% based on online mentions. Models were ran separately for each broad research area and contain the number of articles written in research subfields as controls. Model fit is quantified by the R^2 and number of observations is denoted with N. These results suggest that in most research areas the four variable groups are strong positive correlates of online success as long as we do not control for scholars' gender, i.e., we investigate the male baseline. Specifically, scientific impact is significantly and positively associated with the online success of men in all broad research areas but Computer Science and Mathematical Sciences. Except for Astronomy and Geosciences, social capital has a significant positive association with online success for men and it is the strongest coefficient overall in Mathematical Sciences. Having a high network femaleness is significantly and positively related to the online success of male scholars across the board and it is especially important in Agricultural, Biological, and Computer Sciences. Conversely, network maleness is a significant characteristic in all broad research areas. When we control for author's gender in the model, i.e., we add the interaction term of being female, the relationship between the same variable groups and online success reverses, weakens or becomes non-significant. In particular, scientific impact does *not* retain its significant positive association with online success in Computer Science, Engineering, Mathematical Sciences, Physics, and Social Sciences. This indicates that there is no reliable link between scientific impact and online success for women in these areas. When the link exists, it is weaker for women's productivity and impact than men's in all areas but Agricultural Sciences. Most notably, in the research areas with the highest representation in our sample (i.e., Chemistry, Biological, Medical, and Geosciences, which together make up 60% of the scientists on Altmetric), women have a lower online success for similar levels of scientific impact than men. Additionally, social capital is significantly and negatively associated with online success for female scientists in Agricultural Sciences, Astronomy, and Biological Sciences. Being embedded into highly female scientific networks is associated with lower online success in Biological Sciences, Engineering, Geosciences, and Medical Sciences. Finally, in 10 out of 12 fields female scientists' online success is low if they have highly masculine co-authorship networks. Taken together, while male scholars' online success is linked with their impact, social capital, and gendered collaboration tie formation in various broad research areas, the same characteristics are not clearly associated with the online success of women.

	N	N^*
Number of Web of Science articles published in 2012 with mention(s) on Altmetric	333,917	241,386
Number of unique scholars with articles mentioned on Altmetric	1,101,076	537,486
Number of articles published in 2012 based on WoS	1,823,069	1,042,928
Number of unique scholars who published articles in 2012 based on WoS	820,568	757,527
Number of articles in the Open Academic Graph collaboration network	13,030,628	12,605,249
Number of scholars in the Open Academic Graph collaboration network	8,684,148	7,373,953
Total number of article shares recorded on Altmetric		4,689,423
Number of article shares on Twitter		3,634,714 (77.51%)
Number of article shares on Facebook		473,884 (10.11%)
Number of article shares on news sites		206,456 (4.40%)
Number of article shares on blogs		177,536 (3.79%)
Number of article shares on Google Plus		90,128 (1.92%)
Number of article shares on Wikipedia		73,405 (1.57%)
Number of article shares on video streaming channels (i.e., YouTube)		16,702 (0.36%)
Number of article shares on Reddit		13,963 (0.30%)
Number of article shares on Q&A sites (i.e., Stackoverflow)		2,635 (0.06%)

Table S2. Basic descriptive statistics of the created data set. N is the number of data points in the original dataset, N^* denotes the number of data points after removing articles with more than 10 authors. Our data use the unique document object identifier (DOI) of research articles to combine information from the Web of Science (i.e., meta-data about articles such as authors and research areas) with Altmetric (online mentions) and the Open Academic Graph (publication history). After connecting these three data sources and excluding research articles with more than 10 authors, our data contained detailed information about the attributes of each article, its authors, and their past co-author teams for 241,386 articles of 537,486 scholars in total.

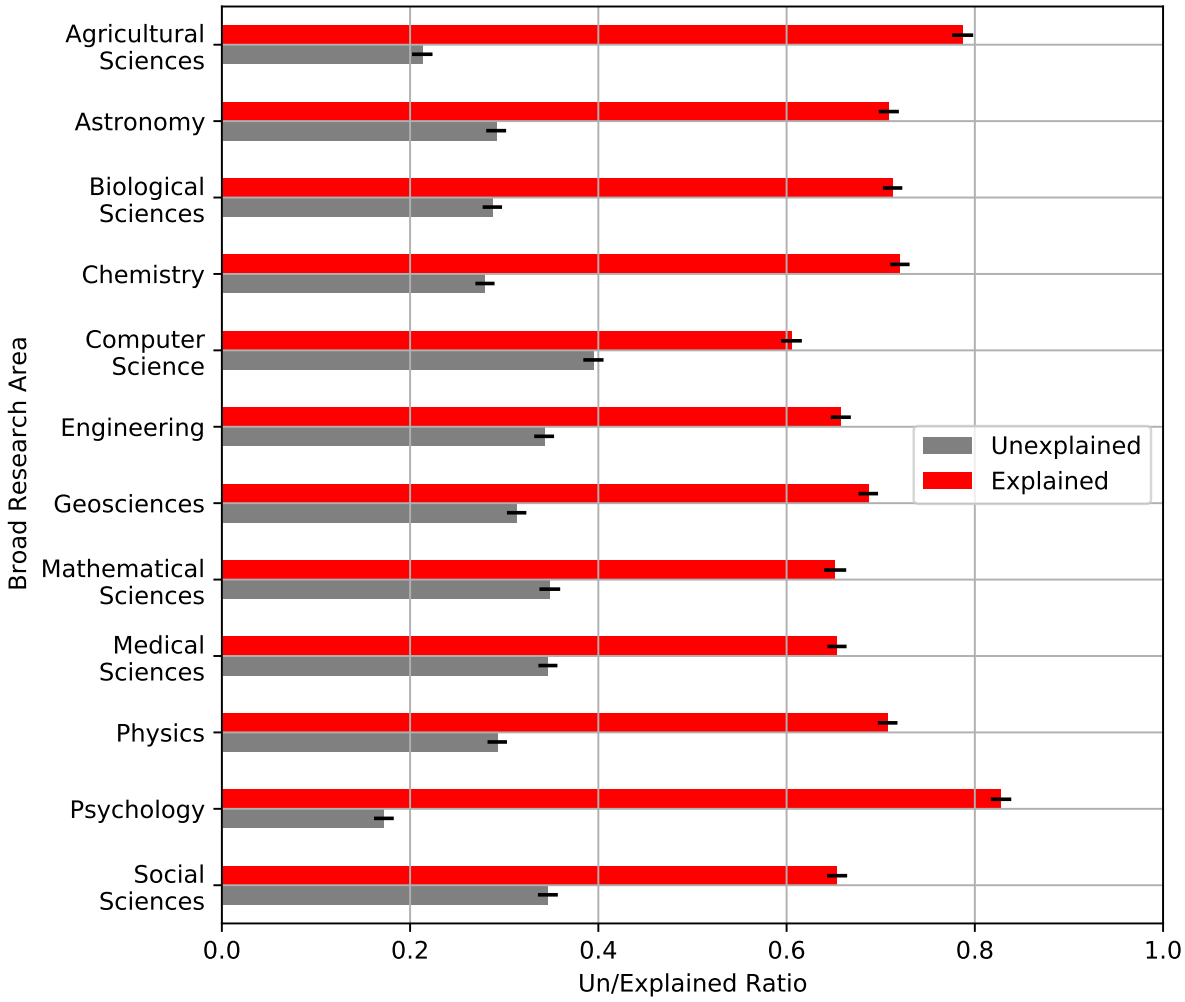

Fig. S3. Explained vs unexplained variance according to a twofold Oaxaca decomposition (1). Across the broad scientific areas, 61-83% of the log number of shares differential can be explained by the effects of differences in the variables we used in our logistic regression models (scientific impact, social capital, network femaleness, network maleness, including controls for the number of articles published per research area). Whiskers indicate 95% confidence intervals.

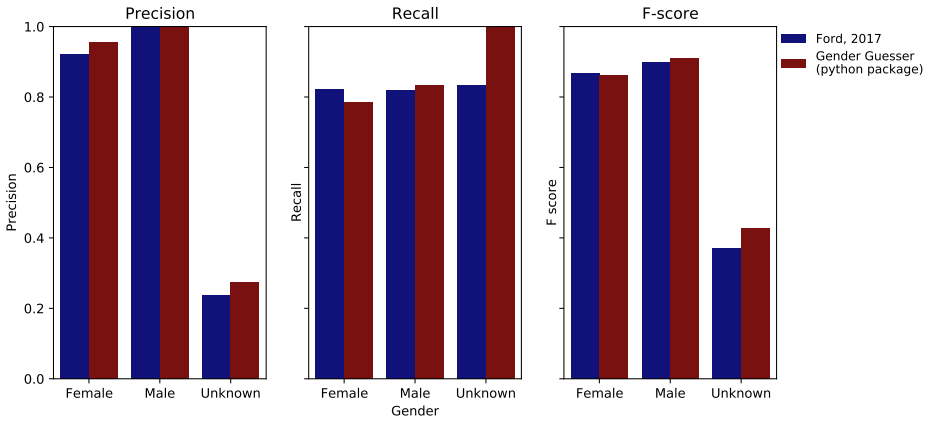

Fig. S4. Accuracy of the used gender imputation Ford et al.'s algorithm and the standard Gender Guesser Python package in comparison with the hand-curated baseline. Precision measures how many scholars in the female, male, and unknown categories were assigned their correct gender label. Recall captures how many of the female, male, and the scholars with unknown gender were correctly identified. The F1-score takes the harmonic average of precision and recall reaching 1 when both metrics are perfect.

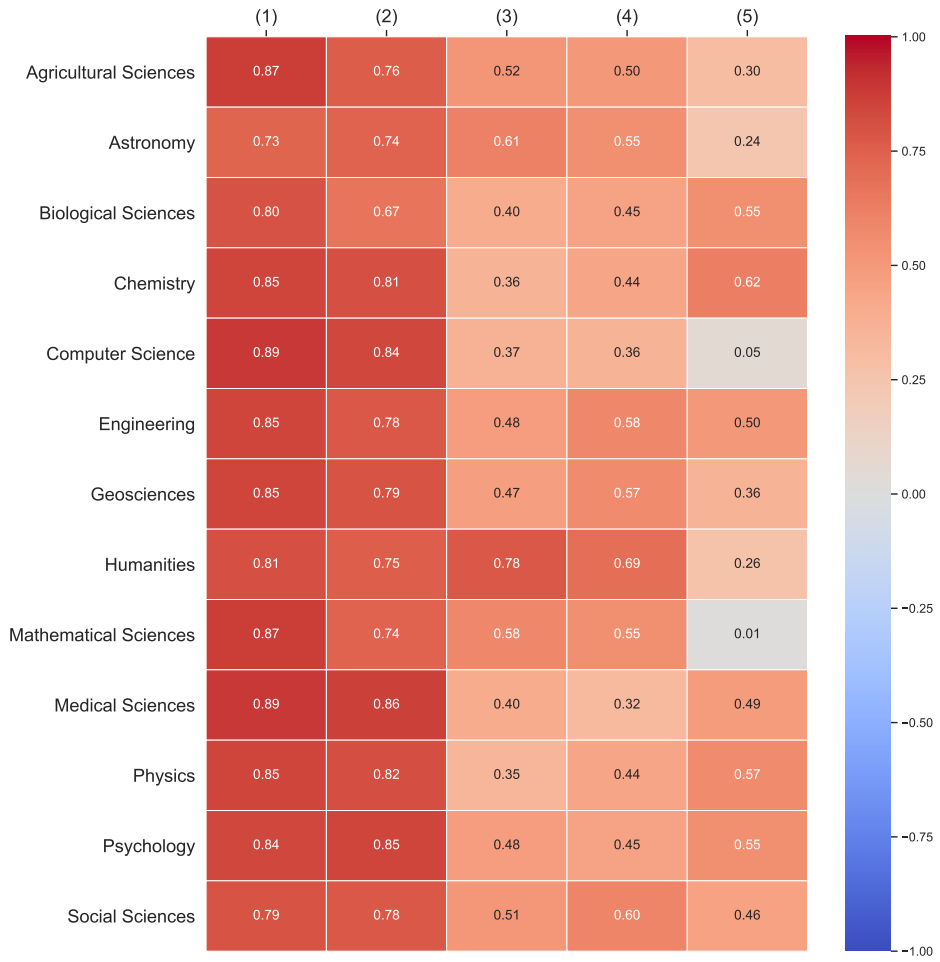

Fig. S5. Pearson correlation between individual variables and the resulting principal component that captures scientific impact. Individual variables: (1) scientific success measured by the *h-index* of scholars in 2012; (2) previous productivity defined as the number of articles researchers wrote in the preceding five years; (3) total *impact factor* of the journals where the articles were published; (4) the number of articles published in *high-impact journals*;(5) and the number of articles on a *hot topic*, which is defined as the top 20% of most shared topics in a broad research area. Correlations are shown for all broad research areas.

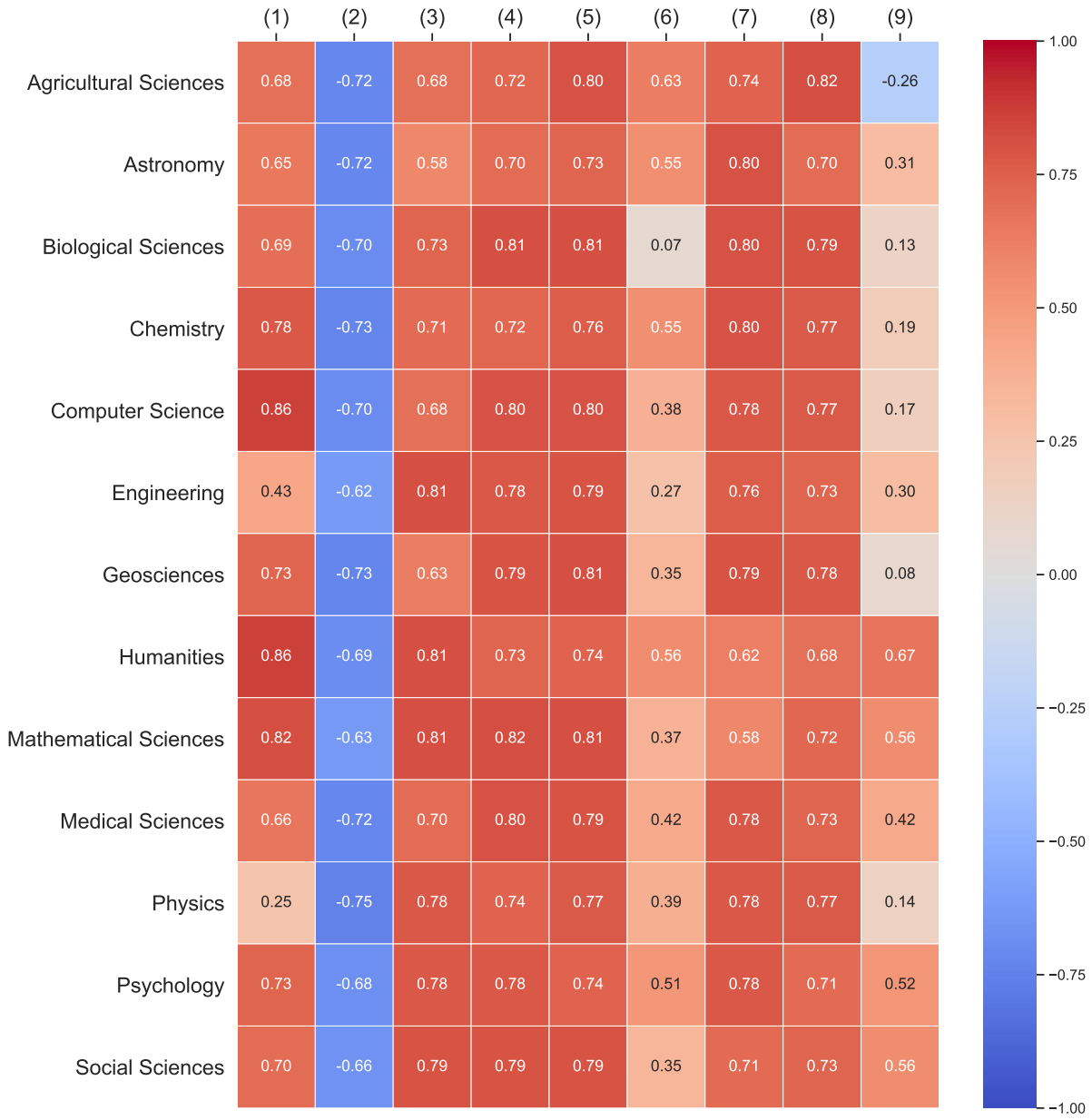

Fig. S6. Pearson correlation between individual variables and the resulting principal components that capture social capital and gendered tie formation. Individual variables: (1) scholars' *number of collaborators* in the previous five years; (2) the *density of their ego network* defined as a sub-network containing scholars, their direct co-authors and all collaborations among those co-authors; (3) the *average size of co-author teams* on individual articles during this time; $(4-5)$ the *number of papers in female/male-majority teams* based on the average female/male ratio in each broad research area; $(6-7)$ the *female/male homophily* among co-authors as the ratio of female-female/male-male ties; and $(8-9)$ the *average tie strength to women/men* which equals the average number of papers co-authored with women/men. Correlations are shown for all broad research areas.

Gender	Altmetric		Web of Science		Open Academic Graph	
	N	%	N	%	N	%
Female	216,646	28.60%	274,681	23.07%	390,891	29.35%
Male	391,013	51.62%	465,185	39.06%	642,507	48.24%
Unknown	149,868	19.78%	451,004	37.87%	298,569	22.41%
Total	757,527		1,190,870		1,331,967	

Table S3. Results of gender imputation with Ford et al.'s algorithm for the three individual data sets.

Broad Research Area	Explained Variance by PCA			
	Scientific Impact	Social Capital	Network Femaleness	Network Maleness
Agricultural sciences	38.74	48.51	51.85	51.85
Astronomy	36.48	42.32	44.34	44.34
Biological sciences	29.51	49.52	43.82	43.82
Chemistry	34.96	54.50	46.68	46.68
Computer sciences	30.01	56.40	47.76	47.76
Engineering	36.00	40.79	43.58	43.58
Geosciences	33.62	48.97	46.68	46.68
Humanities	39.60	62.48	46.72	46.72
Mathematical sciences	39.01	57.34	49.11	49.11
Medical sciences	33.83	48.50	48.07	48.07
Physics	33.89	40.96	42.88	42.88
Psychology	35.98	53.76	47.05	47.05
Social sciences	34.40	51.60	45.97	45.97

Table S4. Explained variance in the four components created with Principal Component Analysis (PCA).

Broad Research Area	Intercept	Scientific Impact	Female: Scientific Impact	Social Capital	Female: Social Capital	Network Femaleness	Female: Network Femaleness	Network Maleness	Female: Network Maleness	R^2	N
Agricultural sciences	0.04	1.5	1.77	1.35	0.78	1.86	0.5	3.63	0.34	0.36	1,458
	(0.00)	(0.00)	(0.01)	0.02	(0.21)	(0.00)	(0.02)	(0.00)	(0.02)		
Astronomy	0.05	1.97	1.21	0.79	0.71	1.14	1.04	1.65	0.72	0.2	2,678
	(0.00)	(0.00)	(0.15)	(0.01)	(0.10)	(0.16)	(0.77)	(0.00)	(0.22)		
Biological sciences	0.07	1.54	1.37	1.05	0.92	1.47	1.02	1.77	1.04	0.23	61,495
	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)	(0.00)	(0.52)	(0.00)	(0.50)		
Chemistry	0.02	1.5	1.32	1.13	1.11	1.29	1.12	1.91	0.95	0.31	17,438
	(0.00)	(0.00)	(0.01)	(0.00)	(0.31)	(0.00)	(0.20)	(0.00)	(0.78)		
Computer science	0.04	1.33	1.48	1.44	0.93	1.7	1.07	2.43	1.35	0.35	1510
	(0.00)	(0.01)	(0.13)	(0.00)	(0.81)	(0.01)	(0.81)	(0.00)	(0.58)		
Engineering	0.03	1.53	1.16	1.32	0.85	1.44	0.86	1.64	1.04	0.29	2,962
	(0.00)	(0.00)	(0.39)	(0.00)	(0.48)	(0.00)	(0.43)	(0.00)	(0.92)		
Geosciences	0.07	1.91	1.39	0.86	0.98	1.41	1.01	1.89	0.75	0.24	8,790
	(0.00)	(0.00)	(0.00)	(0.00)	(0.85)	(0.00)	(0.91)	(0.00)	(0.04)		
Mathematical sciences	0.05	1.18	0.8	1.63	1.56	1.73	0.96	2.19	0.97	0.32	1062
	(0.00)	(0.13)	(0.63)	(0.00)	(0.22)	(0.00)	(0.86)	(0.00)	(0.93)		
Medical sciences	0.07	1.23	1.12	1.09	1.01	1.56	0.94	1.54	0.87	0.24	121,462
	(0.00)	(0.00)	(0.00)	(0.00)	0.48	(0.00)	(0.00)	(0.00)	(0.00)		
Physics	0.05	1.46	1.24	1.16	0.98	1.36	0.97	1.95	0.75	0.23	9,287
	(0.00)	(0.00)	(0.08)	(0.01)	(0.90)	(0.00)	(0.76)	(0.00)	(0.14)		
Psychology	0.2	1.34	1.05	1.38	1.19	1.36	1	1.62	0.82	0.22	10,670
	(0.00)	(0.00)	(0.40)	(0.00)	(0.01)	(0.00)	(0.98)	(0.00)	(0.00)		
Social sciences	0.09	1.51	1.39	1.74	0.86	1.4	0.95	1.35	0.89	0.25	6,424
	(0.00)	(0.00)	(0.00)	(0.00)	(0.17)	(0.00)	(0.54)	(0.00)	(0.33)		

Table S5. Logistic regression model performance with a definition of online success based on the 5% of the most frequently mentioned scholar. The table shows the odds ratio and significance level of variables, model fit (R^2) and the number of observations (N) in each model ran separately by broad research area.

			Top 25%		
Broad Research Area	Recall	Precision	F1	Accuracy	AUC
Agricultural Sciences	0.55	0.76	0.64	0.76	0.79
Astronomy	0.48	0.71	0.57	0.71	0.75
Biological Sciences	0.64	0.75	0.69	0.72	0.78
Chemistry	0.58	0.80	0.67	0.80	0.84
Computer Science	0.37	0.80	0.50	0.80	0.76
Engineering	0.47	0.79	0.59	0.82	0.82
Geosciences	0.57	0.73	0.64	0.71	0.76
Mathematical Sciences	0.50	0.80	0.61	0.78	0.81
Medical Sciences	0.65	0.75	0.70	0.71	0.77
Physics	0.51	0.75	0.60	0.76	0.79
Psychology	0.87	0.75	0.81	0.72	0.77
Social Sciences	0.66	0.71	0.68	0.67	0.74
			Top 5%		
Broad Research Area	Recall	Precision	F1	Accuracy	AUC
Agricultural Sciences	0.38	0.70	0.49	0.93	0.89
Astronomy	0.16	0.64	0.25	0.91	0.79
Biological Sciences	0.25	0.71	0.37	0.90	0.82
Chemistry	0.22	0.64	0.32	0.97	0.88
Computer Science	0.30	0.69	0.42	0.96	0.90
Engineering					
Geosciences	0.24	0.67	0.36	0.90	0.83
Mathematical Sciences	0.30	0.64	0.41	0.92	0.88
Medical Sciences	0.28	0.73	0.41	0.89	0.81
Physics	0.19	0.73	0.30	0.94	0.83
Psychology	0.39	0.72	0.50	0.81	0.80
Social Sciences	0.31	0.71	0.43	0.88	0.82

Table S6. Alternative evaluations of the accuracy of logistic regression models predicting the top 25% and top 5% most successful scientists based on online mentions.

Broad Research Area	Intercept	Scientific Impact	Female: Scientific Impact	Social Capital	Female: Social Capital	Network Femaleness	Female: Network Femaleness	Network Maleness	Female: Network Maleness	R^2	N
Agricultural Sciences	0.43	1.30	1.52	1.87	0.61	1.63	0.76	2.61	0.52	0.22	718
	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.00)	(0.05)	(0.00)	(0.00)		
Astronomy	0.26	1.52	1.11	0.96	0.9	1.21	0.89	1.63	0.75	0.16	808
	(0.00)	(0.00)	(0.37)	(0.45)	(0.42)	(0.00)	(0.20)	(0.00)	(0.04)		
Biological Sciences	0.68	1.73	1.24	1.31	0.9	1.6	0.84	1.83	0.79	0.2	34,390
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)		
Chemistry	0.17	1.43	1.3	1.55	0.87	1.3	0.99	1.77	0.68	0.31	7232
	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)	(0.00)	(0.88)	(0.00)	(0.00)		
Computer Science	0.3	1.1	1	1.45	1.28	1.99	0.75	2.35	0.7	0.19	518
	(0.00	(0.17)	(0.98)	(0.00)	(0.25)	(0.00)	(0.06)	(0.00)	(0.10)		
Engineering	0.26	1.46	1.46	1.78	0.94	1.86	0.79	2.09	0.8	0.27	958
	(0.00)	(0.00)	(0.03)	(0.00)	(0.74)	(0.00	(0.04)	(0.00)	(0.21)		
Geo Sciences	0.60	1.72	1.37	1.03	0.9	1.59	0.84	1.8	0.68	0.17	3,468
	(0.00)	(0.00)	(0.00)	(0.28)	(0.14)	(0.00)	(0.00	(0.00)	(0.00)		
Mathematical Sciences	0.53	0.95	1.13	2.92	0.9	1.57	0.84	1.81	0.62	0.23	278
	(0.00)	(0.56)	(0.71)	(0.00)	(0.74)	(0.00)	(0.35	(0.00)	(0.05)		
Medical Sciences	0.53	1.25	1.11	1.26	0.97	1.56	0.93	1.41	0.87	0.19	78,922
	(0.00)	(0.00)	(0.00)	(0.00)	(0.06)	(0.00)	(0.00)	(0.00	(0.00)		
Physics	0.45	1.58	0.96	1.63	1.18	1.3	1.05	1.96	0.78	0.22	2,430
	(0.00)	(0.00)	(0.68)	(0.00)	(0.10)	(0.00)	(0.37)	(0.00)	(0.00)		
Psychology	1.88	1.26	1.05	1.7	1.05	1.52	0.98	1.59	0.93	0.18	8,330
	(0.00)	(0.00)	(0.45)	(0.00)	(0.49)	(0.00)	(0.70)	(0.00)	(0.24)		
Social Sciences	0.83	1.32	1.17	1.81	0.93	1.37	0.96	1.22	0.91	0.15	3,338
(0.00)	(0.00)	(0.09)	(0.00)	(0.45)	(0.00)	(0.62)	(0.00)	(0.27)			

Table S7. Average model performance based on 5 gender-balanced samples (i.e, samples contained same number of men and women). Success is defined as being in the top 25%. The table shows the odds ratio and significance level of variables, model fit (R^2) and the number of observations (N) in each model ran separately by broad research area.

	Top 25% with balanced sample				
Broad Research Area	Recall	Precision	F1	Accuracy	AUC
Agricultural Sciences	0.54	0.76	0.63	0.76	0.79
Astronomy	0.48	0.71	0.57	0.71	0.75
Biological Sciences	0.64	0.75	0.69	0.72	0.78
Chemistry	0.58	0.79	0.67	0.80	0.84
Computer Science	0.37	0.80	0.50	0.80	0.76
Engineering	0.47	0.79	0.59	0.82	0.82
Geo Sciences	0.56	0.73	0.64	0.71	0.76
Mathematical Sciences	0.50	0.80	0.61	0.78	0.81
Medical Sciences	0.65	0.75	0.70	0.70	0.77
Physics	0.51	0.75	0.60	0.76	0.79
Psychology	0.87	0.75	0.81	0.72	0.77
Social Sciences	0.66	0.71	0.68	0.67	0.74

Table S8. Alternative quantification of the average accuracy of 5 gender-balanced logistic regression models predicting the top 25% most successful scientists.