Science AAAS

www.sciencemag.org/cgi/content/full/science.1227833/DC1

Supplementary Materials for

Flows of Research Manuscripts Among Scientific Journals Reveal Hidden Submission Patterns

V. Calcagno,* E. Demoinet, K. Gollner, L. Guidi, D. Ruths, C. de Mazancourt

*To whom correspondence should be addressed. E-mail: vincent.calcagno@sophia.inra.fr

Published 11 October 2012 on *Science* Express
DOI: 10.1126/science.1227833

This PDF file includes:

Materials and Methods
Figs. S1 and S3 to S5
Tables S1 to S3
References

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/science.1227833/DC1)

Fig. S2
Databases S1 and S2 as Excel files

Materials and Methods

Submission data acquisition

–Survey of corresponding authors

A target group of journals was assembled from subject categories in Thomson ISI Journal Citation Reports (JCR). We selected 16 subject categories, aiming at building a consistent set within feasibility constraints. We also included three major multidisciplinary journals (Nature, Science, and PNAS). Note that subject categories are partly overlapping in terms of journals (see Additional Data table S1 in the SOM). This summed up to 917 unique journals (Table S1). In December 2008, we downloaded for all of them the ISI Web of Science "full record" of every article published between 2006 and 2008, restricting *database* to "Science Citation Expanded", *article type* to "Research article", and *language* to "English". This yielded about 250,000 references. Text files were converted to the BibTeX format and screened to remove possible duplicates or incomplete records (e.g. without an email for correspondence). The BibTeX database was parsed with a custom JAVA application using the *bibtex* package to extract relevant information for each reference. At this stage, if one author (as identified by an email address) was associated with more than 5 articles, 5 articles were selected at random and the others were dropped, to limit the variance in author representation and avoid harassing any individual author with too many questions. This represented less than 1% of all articles; 85% of corresponding authors were sampled only once.

A final list of 215,084 articles was thus obtained and used for the survey. For each article, a personalized email was sent to the corresponding author. Emails were prepared from a template (Supplementary Figure 1) and sent as plain text from a dedicated email box at McGill University (publiweb@mcgill.ca), to minimize the risk of being flagged as spam and maximize the rate of response (*27*). Authors were asked to reply directly to the email with their response (either nothing if the publishing journal was the first choice or the name of the journal previously targeted), following a $ symbol. Note that a dollar symbol was required even when no other journal had been targeted before, so that all types of answers demanded a similar effort of typing something on the keyboard before clicking.

Emails were sent with a JAVA automaton between December 2008 and February 2009. Replies on the email box were collected in August 2009, by downloading the inbox folder as an .mbox file.

–Processing of email replies

About 100,000 emails had been received by August 2009. The .mbox file was read with a custom JAVA application using the *javamail* and *gnumail* packages. Emails were parsed for sender email address, title and body content, and converted to a common format: one single plain-text part with UTF8 character encoding. Each email went through a first scan: title and body content were matched against regular expressions to locate the expected patterns for replies (most importantly, the dollar symbol that had been requested). This filtered out various junk emails, automatic replies or server notifications. The remaining emails were considered as candidate replies and scanned more thoroughly with the Mathematica 8 software (Wolfram Research Inc.). For this purpose, different scenarios corresponding to acceptable replies were implemented as a set of regular expressions and other deterministic rules, and character strings following dollar symbols were extracted from acceptable emails. Note that all answers that did not follow the expected syntax were discarded at this stage. The last step was to interpret the character strings and determine the response, be it "yes, this was the first journal attempted" or "no, we submitted to this other journal before". This was performed in Mathematica with a semi-automatic procedure: a program first tried to interpret the answers using regular expressions and fuzzy matching, by comparing the character sequences and the ISI JCR list of journal names. When the program could not come up with an unambiguous interpretation, it returned a list of suggestions and prompted its human operator for the correct interpretation. Answers were mapped to original queries using the integer tag in the body text (when available) or a combination of the email address and the name of the publishing journal (otherwise). The analysis produced a total of 80,748 usable answers, each describing the submission history of a published article.

–Response rate and assessment of potential bias

The effective response rate was above 30%. Sources of non-response included, besides unwillingness to participate, no-longer valid email addresses, failures to deliver email (due to restrictive spam filters), misdirected responses (some 2,000 responses were sent to VC's box and thus ignored), failure to follow the instructions (e.g. omission of the dollar symbol when replying), incapacity to remember the submission history of an article, and a 48-hour breakdown of the email server in February 2008 caused by excessive email traffic.

The response rate varied across subject categories (Table S1) but in an idiosyncratic way: there was no association with similarity to the investigator's own field of research (ecology/evolutionary biology). Neither did it present any correlation with journal impact factor (Pearson r^2=0.012; n=917 journals). More recent articles were more likely to receive a reply, but this between-years variation was very small compared to the between-fields variation and was similar in all fields (Binomial model; AUROC reduced by 70% if field dropped; 10% if year dropped; 2% if interaction dropped). It was also verified with quantile-quantile plots that the distribution of citation count values among responses was representative of the actual distribution (i.e. the distribution for all articles, including those for which we had no reply). All these diagnostics suggest that response bias was minimal. Nevertheless, we will quantify the potential effect of arbitrary levels of response bias on the estimation of the percentage of first-intents, since this is the result most vulnerable to bias.

Estimation of the overall percentage of first-intents

The percentage of first-intents (fraction of articles initially submitted to the publishing journal) showed some variation across subject categories but, as for the response rate, this variation was not associated with the similarity to the investigator's field of research (Table S1). It was uncorrelated to the variation of response rate (Table S1; Pearson r^2=0.003). The percentage of first-intents declined steadily with publication year, symmetrically in all subject categories, with a difference of 6% between 2006 and 2008. This might result from non-response bias, with greater difficulty to remember non-trivial submission histories for older articles (as some participants reported). Under this interpretation the percentage of first-intents observed for articles published in 2008 would be more accurate, yielding a slightly lower value of 73%. This could still be an overestimate if responding positively were significantly easier than responding negatively. The population surveyed, the non-sensible nature of the question and the survey protocol make it unlikely that the latter effect be strong (*27*).

Given that we got replies from over 37% of the entire population, it is informative to estimate the true percentage of first-intents under arbitrary strengths of the putative non-response bias. Let us write the non-response bias, θ, the probability that an article was a first-intent (FI) among the set of non-answered requests (NR), relative to the observed probability in the set of responses (R), i.e. $\theta=P(FI \mid NR)/P(FI \mid R)$. Under these conditions, the true percentage of first-intents in the whole population of articles is $P(FI \mid R)*[P(R)+\theta*(1-P(R))]$, with $P(R)=37\%$ and $P(FI \mid R)=75\%$ (Table S1). From this we can estimate the true value P(FI), conditional on arbitrary levels of response bias. For example, $\theta=1$ would mean no bias, whereas $\theta=2$ would mean that non-respondents were twice as likely to have published a first-intent as respondents. Even assuming $\theta=0.5$, i.e. a very strong bias against resubmissions, the true percentage of first-intents would still be greater than 50%.

Network analysis

–Graph construction and visualization

All articles that were submitted to another journal before the publishing journal were used to construct a weighted and directed graph (available as additional Data table S2 of the SOM). Each vertex corresponded to a journal (as identified by its official ISI abbreviated name) and each edge from vertex A to vertex B corresponded to a resubmission from journal A to journal B (with eventual publication in journal B). Self-loops represented first-intents. Edges were weighted according to the number of articles that had this final submission history. For visualization (Fig. 1), the graph was treated as a simple unweighted graph and drawn in a plane using the Frichterman Reingold layout, as common in related studies (*1, 10, 24*). The algorithm models journals as physical particles that tend to repel one another but are tied by edges, and iteratively moves them under these constraints. Figures 1 and S2 were prepared with Gephi (*28*).

–Graph analysis and community detection

We analyzed the network with the *igraph* library as implemented in R. The 18 journals that did not belong to the giant connected component were ignored. To evaluate the overall importance of journals in the network, we computed closeness, betweenness and eigenvector centrality metrics (*20*). All had a similar positive association with journal impact factor, and we report closeness centrality (Fig. S4; Pearson r^2=0.25). In the main text we decomposed this association into interpretable components (Figs 2 and 3). We identified clusters of journals with preferential connections (or sub-networks) by maximizing modularity, i.e. the number of edges within clusters relative to among clusters. This was performed with the spinglass algorithm with default parameters as implemented in the igraph library (*19, 29*). As the community detection algorithm is stochastic, we ran 100 replicates analyses. Modularity values all lied in the range (0.506, 0.509), and we retained the partition with highest modularity. Cluster definitions showed slight differences among replicate runs, most variation being about the quality of the discrimination of “Entomology” and “Environmental Sciences” as independent clusters. All reported results were robust to this variability. The seven biggest sub-networks (out of 16) were retained for illustration, based on a size threshold (more than 5% of nodes). To compare the ISI subject categories and the journal clusters derived community detection, the fraction of journals assigned to the different clusters was computed for each subject category, resulting in a 17x7 matrix akin to a contingency table (Fig. S3). For more intuitive visualization, each subject category was also projected on Figure 1 as the centroid of all its journals. Network reciprocity was computed using the reciprocity function in the *igraph* library for R.

–Statistical analyses in relation to impact factor

Impact factors (IF) were log-transformed since on this scale they are approximately normally distributed across journals. Each vertex (journal) was attributed an IF value downloaded from ISI JCR (2008 edition). ISI impact factor is neither the only nor the best index of journal importance (*30*), but was the most widespread and thus the more likely to influence submission decisions, in years 2005-2007. Similar results were obtained using 2006 impact factors, which were strongly correlated with 2008 values (r^2 = 0.7).

To test for the role of journal impact factor in influencing resubmission patterns (Fig. 2B), we generated random graphs by rewiring edges according to specific null hypotheses: (i) the previously targeted journal was picked at random (preserving the in-degree distribution), (ii) the previously targeted journal was picked with odds proportional to the observed out-degree (thus preserving both the in- and out-degree distributions. Note that since no efficient algorithm is available to sample permuted graphs uniformly in this context, the Monte-Carlo approach only preserved the out-degree distribution asymptotically, unlike a permutation test). Each time 10,000 Monte-Carlo replicates were generated to produce the corresponding null distributions of the IF differential (Fig.2B).

The trends with impact factor in Figures 2A (log-transformed counts) and 3 (proportions, weighted according to the number of observations per journal) were modeled with Generalized Additive Models (GAMs) as implemented in the *mgcv* library for R, function *gam* (R Development Core Team, 2011). The optimal complexity of the smooth with respect to impact factor was determined through cross-validation as implemented in the library. Similar trends were obtained with loess smoothing or cubic splines.

For Figure 3, we repeated the rank-correlation tests for each subset of journal corresponding to one subject category (Table S1), to evaluate the consistency of the overall trend. Spearman's rank correlations were negative in all but three categories (Table S2). Given the much reduced statistical power (one journal is one observation in this analysis), some statistical significance was achieved in only four categories (Table S2). All significant correlations were negative.

Citation counts analysis

–Citation data

In July 2011, using the same search criteria as in December 2008, we re-downloaded from ISI Web of Science the full records of all articles we had sampled. They were converted to BibTeX and parsed with a JAVA program as before, this time also extracting the “Times cited” (TC) field. Each article for which an answer had been obtained was then looked for in the 2011 records and assigned its TC value.

–Statistical analysis

To test the effect of submission history on the number of times an article was cited (TC; Fig. 4), we used an exact permutation approach: TC values were randomly permuted with respect to submission histories, within each journal-by-year combination. Indeed, as expected, TC values were significantly affected by the year of publication (earlier articles being more cited as of July 2011 on average) and by the publishing journal (more prestigious, e.g. high impact factor, journals being more cited on average). This permutation procedure thus amounted to including year and journal as fixed categorical effects and controlling for them and their pairwise interaction, in an ANOVA setting. For each of 10,000 permuted data sets, the difference in TC between the submission histories was computed, to generate their null distribution and assess significance. Note that the distribution of TC values was extremely skewed, so the mean could not be safely used as a location parameter for the test. We thus used two test statistics: (i) the difference between the means of the log-transformed TC (TC'=log(TC+1), as TC' values were almost normally distributed), and (ii) Wilcoxon's rank-based difference in TC. The latter approach is more robust since it is invariant to arbitrary monotone transformations of TC, whereas the former returns values that are more interpretable in terms of citation numbers. Both yielded identical conclusions (Table S3). The difference between within-cluster and between-cluster resubmissions was robust to variations in cluster definition by the community-detection algorithm (Table S3). In figure 4B multidisciplinary journals were excluded (for they are not reliably assigned to one thematic cluster) and only the seven biggest clusters were considered. Include all journals and all clusters did not affect the results.

Title: *a question about your article in ZZZZ*

Dear Dr XXXX,
I am conducting a survey of the publication process in biological sciences, and one of your articles has been selected for this survey (see below).

Was ZZZZ the first journal to which you submitted your manuscript "YYYY"?

IF yes, please reply to this email (hit the "reply" button), and in the body of your reply, type the $ (dollar) symbol.
IF no, please reply to this email and in the body of the reply, type the $ symbol followed by the name of the immediately previous journal to which you submitted the manuscript (not the complete sequence of submissions, just the journal immediately before).

NOTA: No need to erase the original message in your reply.

WHO IS ASKING:
Dr Vincent Calcagno, postdoctoral fellow in theoretical ecology at McGill University.
http://redpath-staff.mcgill.ca/calcagno/index.html
My supervisor is Claire de Mazancourt (claire.demazancourt@mcgill.ca).

WHAT FOR:
This is an ongoing research project. We are conducting a survey of the publication process in the years 2006-2008. Our goal is to study the fluxes of articles between scientific journals. Visit http://redpath-staff.mcgill.ca/calcagno/proj.html for details. Be sure that all the data we collect will be analysed anonymously. This project has been approved by the McGill Ethical Board.

We use this email box for the purpose of the survey only, so please don't use it to ask questions and don't attach any file to your reply. If you'd like more information about my research, you can visit http://redpath-staff.mcgill.ca/calcagno/proj.html, where there is a FAQ section, or email me at my regular address vincent.calcagno@mcgill.ca (please do not send your reply to my regular address, since replies will be analysed automatically).

Thank you very much for your cooperation!
Yours sincerely,
Vincent Calcagno

PhD, Postdoctoral Fellow
Redpath Museum
McGill University
Montreal, CANADA
http://redpath-staff.mcgill.ca/calcagno/index.html

T=TTTT

Fig. S1.

Email template used for the survey. XXXX was replaced with the name of the corresponding author (or “colleague” if more than one corresponding author was indicated in the ISI record), ZZZZ was replaced with the name of the publishing journal, YYYY with the title of the article, and TTTT with an integer tag for later reference.

Fig. S2. (separate file: 1227833s4.pdf)
High resolution picture of the submission network (Figure 1) as a separate pdf file.

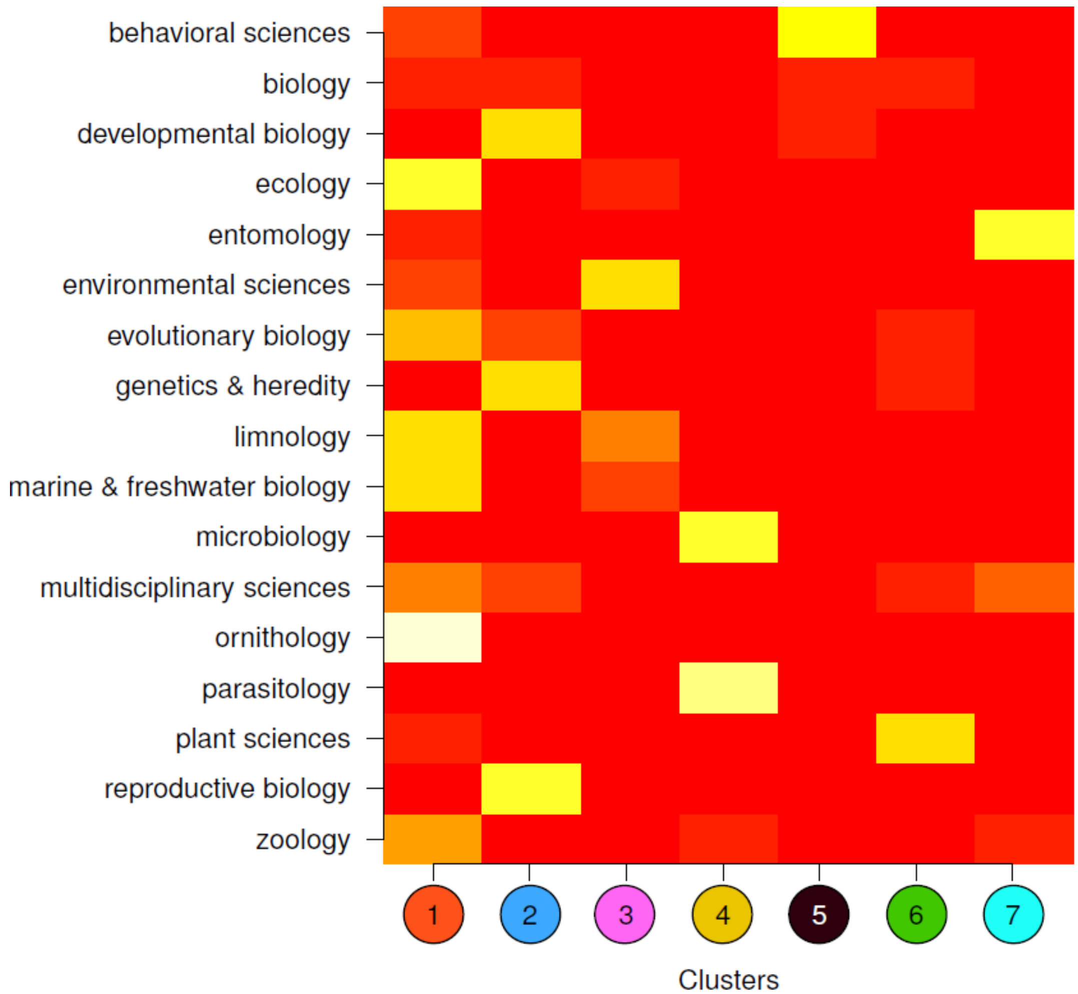
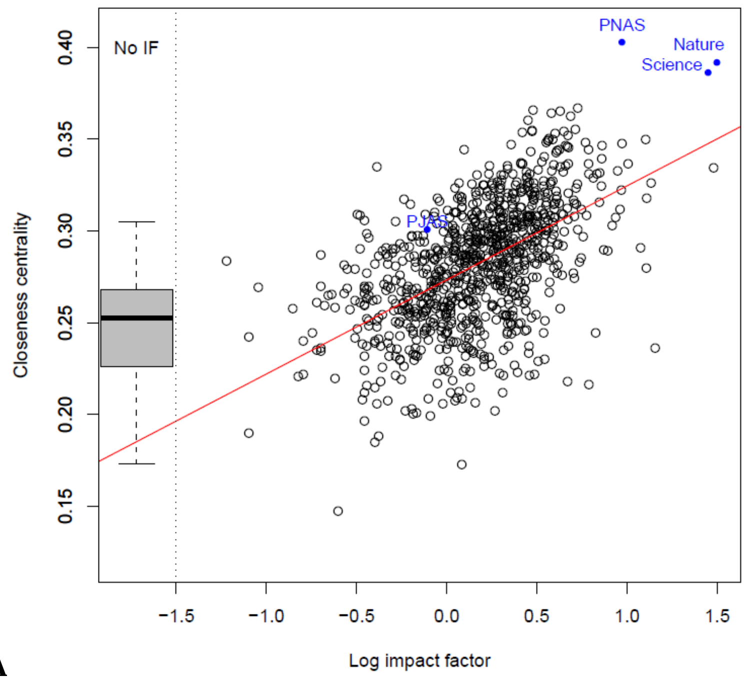
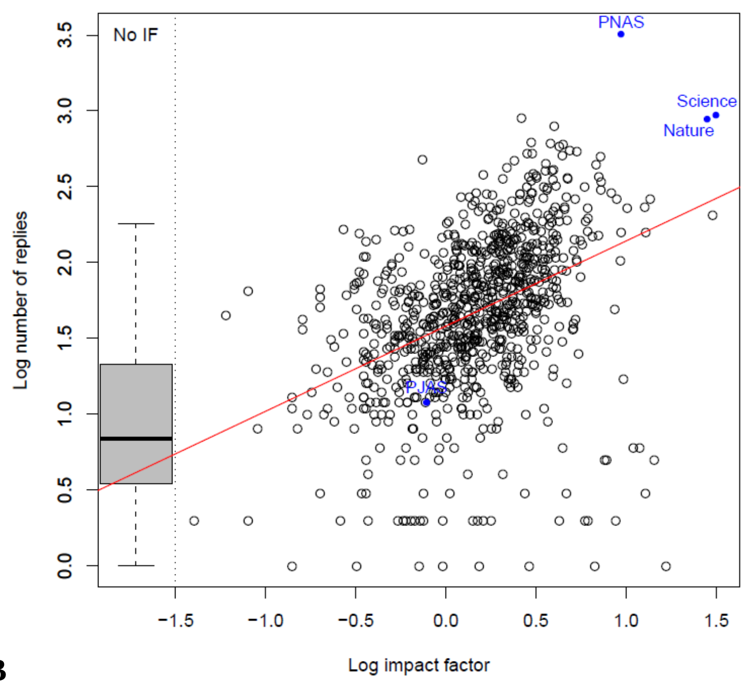


Fig. S3.
The match between ISI subject categories and the major journal clusters derived from network analysis. For each ISI subject category (rows) the fraction of journals assigned to each cluster (columns) is indicated with a color ranging from red (0) to 1 (white), through yellow shades. This shows, for instance, that the "Plant sciences" category is strongly consistent with the sixth cluster.



A



B

Fig. S4.

(A) Journal centrality in the submission network (closeness centrality) in relation to impact factor. (B) Number of replies in relation to impact factor. In both panels, surveyed journals with an ISI impact factor in 2008 (N=911) are shown as a scatter plot and a least-squares linear regression (red line). Journals with no ISI impact factor in 2008 (N=16) are shown as a box-whiskers plot on the left. The four multidisciplinary journals are highlighted in blue.

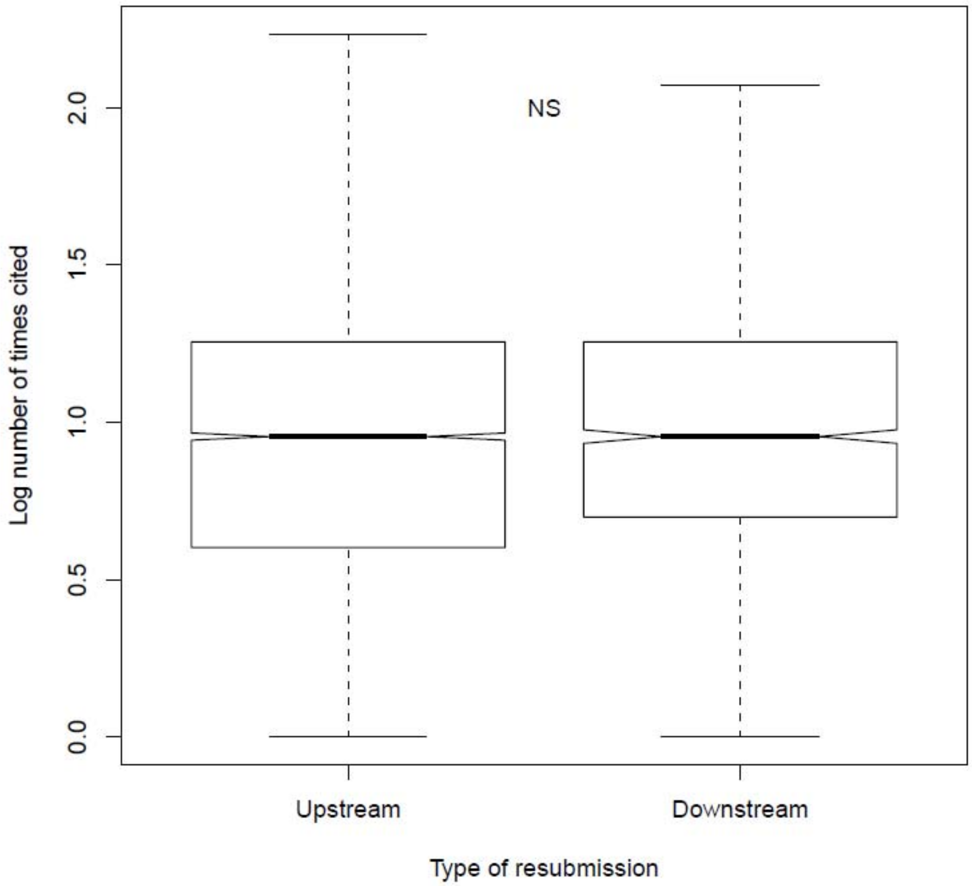


Fig. S5.
Number of times cited for resubmissions to a higher impact journal (left) or lower impact journal (right). N=18,078; same methods as in Figure 4.

Table S1.

The 17 ISI subject categories that were sampled, with the number of journals surveyed, the corresponding number of articles, number of usable replies, effective response rate, and percentage of first-intents. The "Multidisciplinary sciences" category was not surveyed exhaustively: only Nature, Science, PNAS and PJAS were included. ISI subject categories have some overlap (see Table S2), hence the difference between overall numbers and column means/totals.

Subject category	Journals	Articles	Replies	Response rate %	First-intents %
a. Behavioral sciences	41	11501	5057	43.97	73.82
b. Biology	61	13448	5337	39.69	74.72
c. Developmental biology	32	9619	3284	34.14	67.17
d. Ecology	107	31731	15899	50.11	73.8
e. Entomology	62	10480	4675	44.61	84.73
f. Environmental sciences	141	37882	12779	33.73	85.38
g. Evolutionary biology	34	11334	5446	48.05	74.31
h. Genetics & heredity	120	37939	13079	34.47	70.98
i. Limnology	17	3428	1528	44.57	86.91
j. Marine & freshwater biology	72	12741	5151	40.43	80.7
k. Microbiology	51	14612	4247	29.07	76.74
l. Multidisciplinary sciences	*4*	*14535*	*5032*	*34.62*	*69.14*
m. Ornithology	19	1452	685	47.18	74.01
n. Parasitology	21	4066	1112	27.35	76.71
o. Plant sciences	138	20696	6913	33.4	79.36
p. Reproductive biology	23	5706	1523	26.69	78.2
q. Zoology	112	12598	5285	41.95	75.72
OVERALL	**923**	**215084**	**80748**	**37.54**	**76.95**

Table S2.
Correlation between the percentage of first-intent articles published and journal impact factor, by subject category. For each category, the number of usable journals and Spearman's ρ statistic are given. The four categories reaching some statistical significance are highlighted (** $p<0.01$; * $p<0.05$; ° $p<0.1$). The multidisciplinary category was omitted for it was incompletely sampled.

Subject category	Journals	Spearman ρ
a. Behavioral sciences	39	0.05
b. Biology	60	-0.07
c. Developmental biology	**30**	**-0.40***
d. Ecology	103	-0.08
e. Entomology	58	-0.19
f. Environmental sciences	133	-0.10
g. Evolutionary biology	33	-0.22
h. Genetics & heredity	**116**	**-0.24****
i. Limnology	16	0.23
j. Marine & freshwater biology	66	-0.03
k. Microbiology	48	0.09
l. Multidisciplinary sciences	*4*	-
m. Ornithology	19	-0.20
n. Parasitology	21	-0.04
o. Plant sciences	**133**	**-0.14°**
p. Reproductive biology	22	-0.02
q. Zoology	**106**	**-0.15°**

Table S3.
Permutation tests for Figure 4. For each panel and test statistic, the observed value (‘first-intents’ minus ‘resubmissions’ for panel A; ‘between’ minus ‘within’ for panel B), the null distribution (2.5 and 97.5 percentiles) and the corresponding p-value are reported. For (B), the last column reports the percentage of community detection runs that yielded a significant difference in the same direction (%-) versus in the opposite direction (%+; N=100 runs).

Panel	Test-statistic	Observed	Null	P-value	%-/%+
A	Mean log	-0.07408955	-0.06882569 ; -0.06819105	<0.001	_
	Wilcoxon’s W	261797344	269427373 ; 269664186	<0.001	_
B	Mean log	-0.09568382	-0.06572467 ; 0.02768187	<0.001	63/7
	Wilcoxon’s W	313910.5	363849 ; 402770.7	<0.001	70/3

Additional Data table S1 (separate file)

Journal overlap between ISI subject categories. This is a 17*17 matrix as an Excel .xls spreadsheet. Rows and columns correspond to subject categories (indexed as in Table S1). Each row (one subject category), gives the percentage of journals shared with each other category and the percentage of journals unique to the category (on the diagonal; bold face).

Additional Data table S2 (separate file)

The weighted adjacency matrix of the submission network. This is a 2094*2094 matrix stored as an Excel .xlsx spreadsheet. The adjacency matrix is on the first sheet, the names of the nodes on the second. Journals not in the sample set have no self-loops (NA) and zero in-degree.

References

1. J. A. Evans, J. G. Foster, Metaknowledge. *Science* **331**, 721 (2011). doi:10.1126/science.1201765 Medline

2. E. Garfield, Commentary: Fifty years of citation indexing. *Int. J. Epidemiol.* **35**, 1127 (2006). doi:10.1093/ije/dyl190

3. D. J. de Solla Price, Networks of scientific papers. *Science* **149**, 510 (1965). doi:10.1126/science.149.3683.510 Medline

4. K. W. Boyack, R. Klavans, K. Börner, Mapping the backbone of science. *Scientometrics* **64**, 351 (2005). doi:10.1007/s11192-005-0255-6

5. X. Shi, J. Leskovec, D. A. McFarland, Citing for high impact. *Proceedings of the ACM International Conference on Digital Libraries* **2010**, 58 (2010). doi:10.1145/1816123.1816131

6. J. A. Evans, Electronic publication and the narrowing of science and scholarship. *Science* **321**, 395 (2008). doi:10.1126/science.1150473 Medline

7. M. E. J. Newman, Coauthorship networks and patterns of scientific collaboration. *Proc. Natl. Acad. Sci. U.S.A.* **101**, (Suppl 1), 5200 (2004). doi:10.1073/pnas.0307545100 Medline

8. C. L. Giles, I. G. Councill, Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 17599 (2004). doi:10.1073/pnas.0407743101 Medline

9. H. Zuckerman, R. K. Merton, Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva* **9**, 66 (1971). doi:10.1007/BF01553188

10. J. Bollen, H. Van de Sompel, Mapping the structure of science through usage. *Scientometrics* **69**, 227 (2006). doi:10.1007/s11192-006-0151-8

11. J. Bollen, H. Van de Sompel, J. A. Smith, R. Luce, Toward alternative metrics of journal impact: A comparison of download and citation data. *Inf. Process. Manage.* **41**, 1419 (2005). doi:10.1016/j.ipm.2005.03.024

12. F. Rowland, The peer-review process. *Learn. Publ.* **15**, 247 (2002). doi:10.1087/095315102760319206

13. R. Smith, Promoting research into peer review. *BMJ* **309**, 143 (1994). doi:10.1136/bmj.309.6948.143 Medline

14. E. J. Weber, P. P. Katz, J. F. Waeckerle, M. L. Callaham, Author perception of peer review: Impact of review quality and acceptance on satisfaction. *JAMA* **287**, 2790 (2002). doi:10.1001/jama.287.21.2790 Medline

15. P. Clapham, Publish or perish. *Bioscience* **55**, 390 (2005). doi:10.1641/0006-3568(2005)055[0390:POP]2.0.CO;2

16. See Supplementary Materials for methods.

17. E. Garfield, Citation analysis as a tool in journal evaluation. *Science* **178**, 471 (1972). doi:10.1126/science.178.4060.471 Medline

18. S. Saha, S. Saint, D. A. Christakis, Impact factor: A valid measure of journal quality? *J. Med. Libr. Assoc.* **91**, 42 (2003). Medline

19. S. Fortunato, Community detection in graphs. *Phys. Rep.* **486**, 75 (2010). doi:10.1016/j.physrep.2009.11.002

20. S. P. Borgatti, Centrality and network flow. *Soc. Networks* **27**, 55 (2005). doi:10.1016/j.socnet.2004.11.008

21. V. Bakanic, C. McPhail, R. J. Simon, The manuscript review and decision-making process. *Am. Sociol. Rev.* **52**, 631 (1987). doi:10.2307/2095599

22. L. W. Aarssen *et al.*, Bang for your buck: Rejection rates and impact factors in ecological journals. *The Open Ecology Journal* **1**, 14 (2008). doi:10.2174/1874213000801010014

23. P. Cassey, T. M. Blackburn, Publication rejection among ecologists. *Trends Ecol. Evol.* **18**, 375 (2003). doi:10.1016/S0169-5347(03)00160-5

24. J. Bollen *et al.*, Clickstream data yields high-resolution maps of science. *PLoS ONE* **4**, e4803 (2009). doi:10.1371/journal.pone.0004803 Medline

25. How to improve the use of metrics. *Nature* **465**, 870 (2010). doi:10.1038/465870a Medline

26. C. Bergstrom, Eigenfactor: Measuring the value and prestige of scholarly journals. *Coll. Res. Libr. News* **68**, 314 (2007).

27. W. Fan, Z. Yan, Factors affecting response rates of the web survey: A systematic review. *Comput. Human Behav.* **26**, 132 (2010). doi:10.1016/j.chb.2009.10.015

28. M. Bastian, S. Heymann, M. Jacomy, *International AAAI Conference on Weblogs and Social Media* (San Jose, CA, 17–20 May 2009).

29. M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **69**, 026113 (2004). doi:10.1103/PhysRevE.69.026113 Medline

30. J. Bollen, H. Van de Sompel, A. Hagberg, R. Chute, A principal component analysis of 39 scientific impact measures. *PLoS ONE* **4**, e6022 (2009). doi:10.1371/journal.pone.0006022 Medline