Additional methods (not peer reviewed) supporting those reported in "Reputation shortcoming in academic publishing"

Rémi Neveu, PhD; André Neveu, PhD

Note: References to figures labeled as figure XX relate to figure of the article. Those labeled as supplementary figures XX refer to the supplementary materials of the article. Those labeled additional figures XX refer to figures referenced in this file. References to paragraphs in the format §X.X.X. with numbers higher or equal than 1.3.X such as §1.6.6.1 relate to paragraphs of the supplementary materials of the article. References to paragraphs §1.1.X or §1.2.X relate to paragraphs of this file unless explicitly mentioned.

Table of contents

1. Additional methods 2
1.1. Inclusion criteria: additional considerations for editors 2
1.2. Data preprocessing and computation 2
1.2.1. Shared step for the extraction of article data for Nature journals and editors 2
1.2.2. Extraction of article data for Nature journals 3
1.2.3. Extraction of article data for editors 3
1.2.4. Building groups of articles 5
1.3. Control analyses 8
1.3.1. Analysis 1 8
1.3.2. Analysis 2 8
1.3.3. Analysis 3 10
2. Additional figure 1 12

1. Additional methods

1.1.Inclusion criteria: additional considerations for editors

The editorial direction of each journal (i.e. the editor in chief for each specialized *Nature* journal and the editor in chief with his/her head of each section for the two generalist journals) was included together with the senior and associate editors who traditionally handle the submitted manuscripts. The rationale is that the editorial direction may intervene in the decisions made by the editors handling submitted manuscripts and they may handle manuscripts themselves.

A specific consideration applied to *Nature*. Besides the two main sections in which original articles were published (i.e. physical sciences and biological, social and clinical sciences), Nature had editors (n=44) dedicated to sections which included articles without any original data but with "minimal new supporting research findings". For a completeness of the dataset, we therefore included in our analyses the Nature editors of the sections labeled "Opinions" and "News & views" (n=8, 17% of the total sample): these sections seemed to allow the publication of new results, even though more limited, like in original articles. That 6 out of 8 editors of these two sections had a doctorate while only 7 out of 36 of the editors of the other sections (i.e. "Careers", "News & features" and "Multimedia") had a doctorate supports the exclusion of the editors of other sections of *Nature*. Note that including these additional 8 editors for the journal *Nature* did not likely bias the results: only one of these 8 editors was included in the analysis of the publications of the editors' former co-authors (see §1.9.3.1 and §1.9.3.6). This editor had a research experience after the doctorate likewise most of the other editors and represented 1.7% of the analyzed sample of the analysis of §1.9.3.1. Moreover, these 8 editors may contribute to the discussion about the submitted manuscripts during the weekly editorial meetings.

1.2.Data preprocessing and computation

1.2.1. Shared step for the extraction of article data for Nature journals and editors

The additional figure 1 sums up the computational steps and their outcomes involved in the extraction of article data for *Nature* journals and editors. The extraction of article data for *Nature* journals and for *Nature* journals' editors was performed in two steps. The first step

was shared between data extraction for *Nature* journals and data extraction for *Nature* journals' editors. The following step was specific to each data extraction.

The output of this first step led to two distinct datasets (additional figure 1): one with articles published in *Nature* journals in which results were sorted by journal and one with articles published by editors in which results were sorted by editor.

The output of this first step led thus to two distinct datasets (additional figure 1): one with articles published in *Nature* journals in which results were sorted by journal and one with articles published by editors in which results were sorted by editor.

1.2.2. Extraction of article data for Nature journals

First we sorted, by authors, the dataset of articles published in *Nature* journals coming from the output of the first step. This sorting was performed in each journal separately. Thus, for a given article with several authors, this article was referenced in this sorting as many times as it had authors: once for each author of the article. These results allowed us to check, in the third step of the extraction of article data for *Nature* journals, whether an author of an article (i.e. article 1) was also author of another article (i.e. article 2) published in one of the *Nature* journals included in the study before that article (i.e. article 1).

Then, we combined these results (step 2 in additional figure 1) with the output of the first step to obtain the data to answer the question of the evolution across time in *Nature* journals of the percentage of original articles with only new authors and of the percentage of original articles with known intermediate authors.

1.2.3. Extraction of article data for editors

First, we replaced dates of the start of the doctorate (PhD or MD), of the end of the doctorate and of the end of the academic research experience by an adapted average value for editors for whom dates were missing (see §1.6.5). We also replaced the collected editor's affiliations by a subset of words extracted from their affiliation. This subset removed all words that were non specific to the editor's affiliation (e.g. "University of Princeton" was replaced by "Princeton"). The rationale for using this subset of words in the affiliation was that the same author's full affiliation may have been referenced in different ways across articles (e.g. the

university of Princeton may be referenced as “University of Princeton” or “Princeton University”). These differences would have obviously tricked algorithms comparing affiliations.

Then, we sorted, by editors, the dataset of articles published by editors (i.e. during their time spent as researchers before being appointed by *Nature* journals) coming from the output of the first step. We computed for each of these articles its type (original article/review or meta-analysis/other articles) and the editor’s place as author in these articles (i.e. 1st author, 2nd author, last author or any other authorship position). For each editor, we removed all editor’s articles published before the start of the editor’s doctorate (PhD or MD) and after the end of the editor’s last position in academic research because these articles may have higher chances to have been written by homonyms of the editor. Finally, we computed, for each editor, the number of articles, irrespective of their type, which were published by the editor during his/her doctorate (PhD or MD). Because we collected only the years of the start and end of the doctorate and of the end of the last editor’s position in academic research, we defined:

- January 1st for the start of the editor’s doctorate
- December 31st for the end of the editor’s doctorate and for the end of the editor’s last position in academic research.

For each editor, we built a dataset of all authors of articles selected by this process. This dataset contained thus all editor’s co-authors when the editor was working as a researcher.

Additionally, we grouped the articles selected by this process in order to address potential homonyms of the editor’s name (see §1.6.6.1).

Second, we identified editors who had homonyms and removed them from all subsequent analyses (see §1.6.6.2). Then, for each editor, we used the dataset of articles published in *Nature* journals coming from the output of the first step in order to compute in each *Nature* journal included in the study (table 1 of the article):

- the number of articles, of any kind, in which at least one author had one of the editor’s affiliations when this latter was working as a researcher (including the doctorate),
- the number of original articles published by each author irrespective of the authorship position of the author on the article.

Note that these two computations were independent: the second computation was not restricted to authors with one of the editor's former academic research affiliations. The first computation was designed to investigate the number of articles published by authors with one of the former academic research affiliations of the editor. The second computation was designed to investigate the number of articles published by editor's co-authors. These two computations considered only articles published during the two years before the editor's appointment at his/her current *Nature* journal (see §1.6.6.3). We used a duration of two years for consistency considerations with the two years duration used to compute the impact factor. The same two computations were also performed for the two years after the editor's appointment at his/her current *Nature* journal (see §1.6.6.3).

In the second computation, we split authors in two groups: authors who were co-authors of the editor (identified in the second step of the extraction of article data for editors) and the other authors.

In order to check the robustness of the results regarding homonyms, we performed the same computations for editor's co-authors but we restricted editor's co-authors to the first group of authors identified with the algorithm grouping editor's co-authors (see §1.6.6.1).

1.2.4. Building groups of articles

The algorithm we designed split all articles of a given editor into groups of articles so that, in a given group of articles, any combination of two articles shared at least more than the editor's name. For instance, two articles of a given group had in common an author distinct from the editor's one while two others of that group shared the same affiliation for the editor.

The algorithm was initialized as follow: for a given editor, the first article encountered in which one of the affiliations of the authors matched one of the editor collected from the editor's biography was assigned to the first group of articles. In parallel, a first group of keywords was built with the keywords of that article.

Then, in the first step, the algorithm looped over all articles published by the editor except the one identified in the initialization. Each article was compared to all articles of the first group. If one of the following conditions was met:

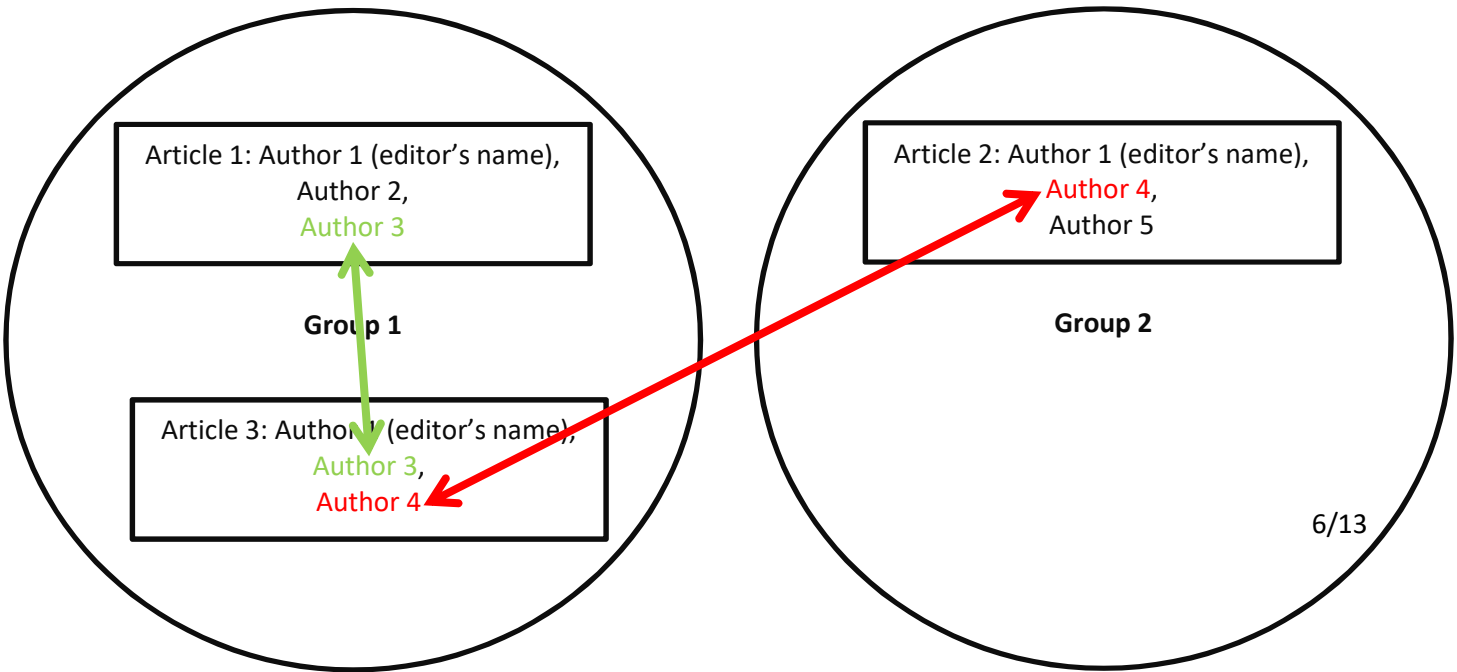
1- the affiliation of the author with the editor's name on the article was one of the affiliations extracted from the editor's biography,

2- the keywords of the article were found in the first group of keywords,
3- the full name of one author of the article (distinct from the editor's name) was the same as the one of an author of one of the articles of the first group,

the article was added to the first group. Its keywords were added to the first group of keywords.

If none of these three conditions was met, the affiliation of the editor was screened for all other groups of articles. If the search was unsuccessful, the same search was performed with keywords (same criterion as in condition 2 mentioned above). If the search was unsuccessful, a final search was performed with authors' names (same criterion as in condition 3 mentioned above). If the search was still unsuccessful, it meant that the author with the editor's name did not share any of the editor's affiliations (either extracted from the editor's biography (i.e. first group of articles) or of the other groups of articles), did not focus on the same research topics as the editor's ones and did not share any of the co-authors. Therefore a new group of articles was created with that article and its keywords built a new group of keywords associated with that group of articles. This new group was added to the groups screened for matching. However, if a group was found (i.e. distinct from the first group), we followed the same procedure as for a successful condition match for the first group: the article and its keywords were merged to that group and its associated group of keywords.

At the end of the first step, all articles of a given editor were therefore classified into groups in which at least more than one author with the editor's name was shared among all article of a given group. However, these clusters may not have been independent. Some groups may still have shared authors because of the order of screening of the articles which cannot be controlled. The following sets illustrate this case:



Article 1 initialized the first group. The second article screened (i.e. article 2) did not match any of the conditions to be included in group 1 (article 3 has not been screened yet and is thus not associated to group 1 yet). It therefore initiated group 2. The third article screened (i.e. article 3) was first checked for a match with group 1 (checks were done by order of creation of groups): this check was successful (because article 1 and article 3 had author 3 in common). It therefore was included in group 1. However article 3 had also an author in common with article 2 which was in group 2 (i.e. author 4 was shared between article 3 and article 2).

The second step of the algorithm aimed thus at merging groups which may have had shared authors' names (distinct from the editor's one). We therefore set group 1 as the reference group of articles. Then all authors (with a name distinct from the editor's one) of the other groups were compared to the authors of the first group. If one of them was found in both groups, then all articles of the group were included in the first group. This comparison between the updated first group and the remaining groups, and the merge with the first group of a group with a successful match, were repeated until no more match was found.

We repeated this procedure for the remaining groups: we defined as a new reference group one of the remaining groups of articles and repeated the computations of the previous paragraph with this new reference group and all remaining groups (i.e. all groups except the first group at the end of the computations of the previous paragraph). These operations were repeated until no more groups could have been compared to the new reference group.

The second step of the algorithm allowed thus to merge groups whose aggregation would have failed in the first step. At the end of the processing, we obtained a first group of articles matched, without biases, by authors' names, keywords and editor's affiliation matching the ones of the editor's biography. Because of its characteristics, this first group may be considered as the one with the lowest probability to contain articles with homonyms of the editor.

This algorithm however may fail for instance to group articles of an editor from the period of his/her doctorate (i.e. period 1) with the articles from the research experience occurring

immediately after the doctorate (i.e. period 2). The failure occurs if there is no shared affiliation among authors of articles of period 1 and of period 2 nor any author (besides the editor) shared between the articles of period 1 and of period 2. The ORCID identifier might have been a solution to this issue but could not be used here because of its late introduction (i.e. the ORCID identifier was set in 2012 while several editors were already working for *Nature* journals).

1.3.Control analyses

1.3.1. Analysis 1

This analysis justified discarding all articles extracted from PubMed and published in the journal *Nature* before 1990.

We used the output of the first step of the extraction of article data for *Nature* journals (see §1.6.1). Over the period 1945-1989, we computed for each year $\left| N_{articles\ published\ each\ year} - average\ of\ N_{articles\ published\ each\ year} \right|$ with $N_{articles\ published\ each\ year}$ being the number of articles, irrespective of their type, published during the year in *Nature*. The average of $N_{articles\ published\ each\ year}$ was computed over all $N_{articles\ published\ each\ year}$ of the period 1945-1989. We refer to this computation as the variations in the yearly number of articles. We did the same for the ratio of the number of original articles to the number of articles, irrespective of their type, published each year in *Nature* (we refer to this computation as the variations in the yearly ratio of original articles). We did the same two computations for the period 1990-2020.

We then performed a two-sample t-test to compare the variations in the number of articles between the two periods. We did the same for the variations in the ratio of original articles.

Finally, we ran a GLM for each period separately. In this GLM, we modeled the variations in the yearly ratio of original articles as a function of year. An absence of association between time and the variations in the yearly ratio of original articles was interpreted as a stable referencing of original articles and a stable space provided to original articles in *Nature*.

1.3.2. Analysis 2

This analysis aimed at checking that, overall, the authors of *Nature* journals' articles were not sorted by alphabetical order of family name. When authors are not sorted by alphabetical order, the ranking, or position, of authors on an article follows the rule:

- the first author has usually contributed most in the study,
- the last author has supervised the work and has a global view of the topic,
- the intermediate authors have provided the remaining specific contributions, their rank, or position, among the other authors of the article matching their contribution to the article by decreasing order of contribution.

We used the output of the first step of the extraction of article data for *Nature* journal (see §1.6.1). For each *Nature* journal, we computed, for each number of authors per article (i.e. articles with 2 authors, articles with 3 authors…), the number of articles in which the authors' names were sorted by alphabetical order. For each number of authors per article, we also computed the total number of articles with this number of authors. In practice, for all articles which had two authors in a given journal, we computed the number of articles whose authors were sorted by alphabetical order in that journal and the total number of articles which had two authors in that journal. We did the same for articles with three authors and so on. Articles with one author were excluded from the analysis for obvious reasons of irrelevance. Note that this computation did not discriminate original articles from the others. We then computed within each journal and within each number of authors in articles the ratio, in %, between the number of articles sorted by alphabetical order and the total number of articles. This ratio is referred to the "observed ratio" thereafter.

Finally we compared this ratio to the probability (in %) to get by chance all authors sorted by alphabetical order in a given article. For n authors on an article, this probability is equal to:

$$p_{alphabetical\ order\ by\ chance} = \frac{1}{n!}$$

We ran a linear mixed model to perform a Bland and Altman analysis. We modeled $2 * \frac{Observed\ ratio - p_{alphabetical\ order\ by\ chance} * 100}{Observed\ ratio + p_{alphabetical\ order\ by\ chance} * 100}$ as a linear function of the number of authors. Journals were set as a random factor over the intercept. In each *Nature* journal, we discarded from the analysis any number of authors for which there was no article in that journal with that number of authors.

If the intercept and the association with the number of authors in that linear mixed model were negative, the probability of having authors sorted by alphabetical order by chance was thus higher than what was observed in *Nature* journals. Such a result was interpreted as an absence of sorting of authors by alphabetical order in articles of *Nature* journals. This justified our discrimination of first, last and intermediate authors in our analyses (see §1.9.3.4 and §1.9.3.5).

We did the same computations and Bland and Altman analysis using original articles only.

1.3.3. Analysis 3

This analysis aimed at checking whether the collection of job offers might have biased the results because of missed job offers: job offers, whose duration for application (which is equal to the duration of publication on the *Nature* website) was shorter than the time elapsed between two consecutive data collection time points, might have been missed.

Using a two sample t-test, we first compared the time elapsed between two consecutive data collection time points to the duration for application to the job offers collected.

Second we quantified the number of job offers which might have been missed because of the time elapsed between two data collection time points of job offers.

We hypothesized that the probability density function of a job offer to be released at a specific date with a given duration for application was uniform across time. The duration of application is equal to the duration of publication of the job offer on the *Nature* website. Therefore, the probability p_{ij} of a job offer to be collected by a given data collection time point, separated from the following time point by d_j, and with a given duration of application d_i' is equal to:

$$p_{ij} = \frac{d_i'}{d_j} \quad \text{if } d_i' < d_j \text{ otherwise } p_{ij}=1$$

Computed over k data collection time points distributed over 365 days, the probability computed over the year becomes:

$$p_i = \sum_{j=1}^{j=k} p_{ij} * \frac{d_j}{365}$$

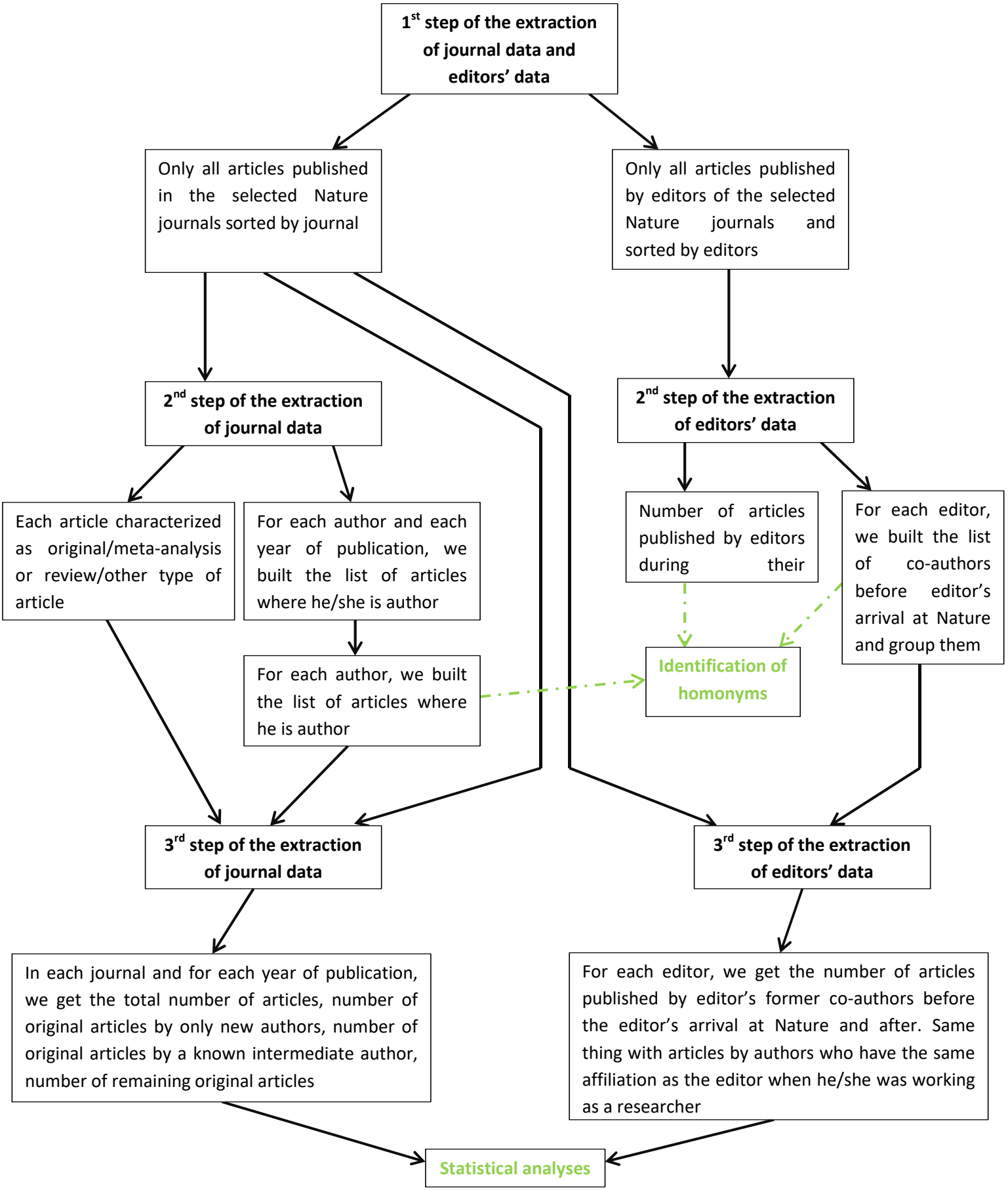
Because the mathematical expectancy of job offers collected for the duration of application d_j equals the number n_i of job offers collected for that duration, we get an estimate of the number N_i of job offers for a duration d_i' which has been released over the year of data collection (i.e. December 2020 to December 2021):

$$N_i = \frac{n_i}{p_i}$$

By summing over the m durations of applications collected, we get an estimate of the total number N of job offers which have been posted by *Springer Nature* between December 2020 and December 2021:

$$N = \sum_{i=1}^{i=m} N_i$$

2. Additional figure 1



Additional figure 1: flow chart of the steps (and their outputs) of the extraction of article data.