Supplementary Information

Career on the move: Geography, Stratification, and Scientific Impact

Pierre Deville, Dashun Wang, Roberta Sinatra, Chaoming Song, Vincent D. Blondel, and Albert-László Barabási

Contents

S1. PREPROCESSING THE DATA

The data provided by the American Physical Society (APS) contains over 450,000 publications, each identified with a unique number and corresponding to all papers published in 9 different journals, namely Physical Review A, B, C, D, E, I, L, ST and Review of Modern Physics, spanning a period of 117 years from 1893 to 2010. For each paper the dataset includes title, date of publication (day,month,year), author names and affiliations of each of the authors. A separate dataset also provides list of citations, using unique paper identifiers.

To obtain accurate individual mobility information of scientists, we select only papers which meet two criteria: (i) no ambiguity between an author and its affiliation is present (an ambiguity is present when more than one author and more than one affiliation are given without any link between them); (ii) having no author group larger than 10. The criterion (i) is necessary to associate at least one affiliation per author, which is an important step to derive individual mobility information, while (ii) is required to identify papers where each author gave a substantial contribution to it. The threshold of 10 authors is chosen by inspection of the distribution of number of authors per paper (Suppl. Fig. S1). The probability density function seems to roughly follow a power-law which is coherent with previous studies [S1]. We observe a deviation from the power-law for papers containing more than 10 authors, suggesting different retribution forms for these large coauthorships. The application of these two filters gives us a final set of 425,369 publications.

S2. AUTHOR NAME DISAMBIGUATION

To derive individual information, one has to reconnect papers belonging to a single scientist. Since no unique author identifier is present in the data, author names must be disambiguated. Grouping authors by their full name is not a solution. Indeed, first names and given names are very often missing or incomplete in the data. As a consequence, author names corresponding to a single scientist can be numerous. Here, we describe a disambiguation procedure that overcomes these inconsistencies by using information about the author but also metadata about the paper.

First, for each author name on a paper, we build an author-paper pair which uniquely identifies the author and his publication. The dataset contains about 1,2 millions of author-

paper pairs and we initially consider each pair to correspond to a unique author.

The principle of the disambiguation process is to merge repeatedly groups of similar author-paper pairs until no further merging is possible. Initially, each group contains a unique author-paper pair. Then, two groups of author-paper pairs are iteratively selected and merged if the three following similarity conditions are fulfilled:

1. Last names in both groups are identical.

2. The first and given names, when available, are compatible.

3. The two groups are citing each other at least once (self-citations) or the two groups share at least one similar co-author or similar affiliations (cosine similarity and tf-idf measure).

The process stops when no more merging is possible. A total of $237,038$ groups and thus distinct scientists are found after the disambiguation process. This represents a decrease of 80% of the number of unique authors initially considered in the dataset.

To validate our approach, we randomly selected from the output 400 pairs of authors with similar names, out if which 200 are considered as a same person by the algorithm and the other 200 are not. We then performed a blind search using Google Scholar to determine, for each pair of authors, if they are indeed the same person or not. We find the false positive rate to be 2% (i.e. authors that are considered as a same person while in reality they are not) and a false negative rate of 12% (i.e. authors that are wrongly categorised as different persons).

S3. AFFILIATION DISAMBIGUATION

For this project, we are interested in the mobility of scientists in term of institutions, i.e. universities or independent research centers. Choosing cities as a proxy for institutions is not a solution here, even though disambiguation would be easier since city names show little variations and are well structured in the data. Indeed, many institutions can be found in a city. The Boston metropolitan area, for example, contains 58 higher educational institutions, 9 of them being major universities [S2]. Aggregating institutions to their city or even zip codes would thus lead to less accurate career paths and would take into account the individual performance of institutions.

A major disadvantage when dealing which such datasets is the inconsistency and errors associated with affiliation names on papers. A total of $319,829$ different affiliation names are identified in the dataset. As an example, $1,655$ of them are found to be associated to the MIT. This wide range of names results from different departments and sub-departments names within institutes, historical changes, abbreviations but also many typographical errors. A procedure resolving these inconsistencies is thus designed in order to disambiguate them.

The disambiguation process is divided in three steps:

1. We geocode all affiliation names present in the data and group them accordingly by coordinates, resulting in $64,107$ geo-tagged groups. This first step dramatically decreases the number of comparisons between elements for the next steps.

2. We compare all affiliation names within a group to each other by using a standard *tf-idf* statistic and cosine similarity measure to merge them into sub-groups. By choosing a high threshold (0.9) for the similarity measure, we ensure that a subgroup corresponds to a single institution.

3. Subgroups are then compared to each other and merged accordingly by using a lower similarity value threshold but also by using the author names previously disambiguated. For each scientist, we compare his affiliations to each other and merge two corresponding subgroups into one if the similarity is high enough (0.7). This particular approach is a key in our procedure. Indeed, similar affiliations on papers authored by a same scientist are likely to correspond to a same institution. Not only it gives us much better results but it speeds up the algorithm by reducing the number of comparison as well.

A total of of $4,052$ groups, i.e. distinct institutions, have been detected in the end.

To validate our results, we use a similar procedure than the one for author names. We randomly select two lists of affiliation pairs: (i) 200 pairs of affiliations that are considered as a single institution and (ii) 200 pairs of affiliations located in the same city but considered as different by the algorithm. We then perform a search using publicly available information to determine, for each pair of affiliations, if they indeed correspond to similar institutions or not. We find the false positive rate to be 11% (i.e. affiliations that are considered as a single institution while in reality they are not) and a false negative rate of 6% (i.e. affiliations that are wrongly categorised as distinct institutions).

S4. DEFINITION OF σ_c

In order to capture the statistical difference in the average citations between papers published before and after the movement of a scientist, one has to normalize this difference by the random expectation when the same scientist's publications are shuffled. Let $c^- = \{c_1^-, c_2^-, ..., c_n^-\}$ and $c^+ = \{c_1^+, c_2^+, ..., c_m^+\}$ be the lists of number of citations for papers published before (c^-) and after (c^+) the transition from i to j ($T_{i,j}$) associated to a scientist. To quantify the change in performance, we introduce

$$\Delta c^* = \frac{\overline{c^+} - \overline{c^-}}{\sigma_c} \tag{S1}$$

where $\overline{c^+}$ and $\overline{c^-}$ are the average of c^+ and c^-, respectively. The quantity σ_c corresponds to the standard deviation of the concatenation of both c^+ and c^- while preserving the moment when the movement took place and is defined by

$$\sigma_c = \sqrt{E[c^2] - E[c]^2} \times \sqrt{\frac{n+m}{n*m}} \tag{S2}$$

where c is the concatenation of both c^+ and c^-, $E[X]$ is the expectation of X and m and n are the size of c^+ and c^-, respectively. The first term in S2 corresponds to the standard deviation of the concatenation of c^+ and c^-. The second term takes into account the size of both lists, setting the number of publications before and after the transition as it is important to compute the random expectation.

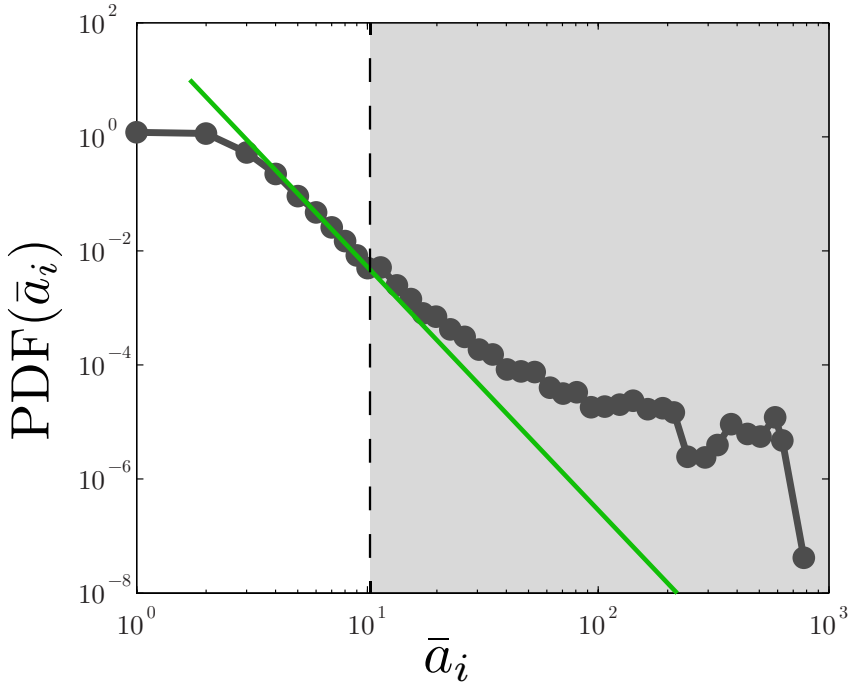

Figure S1: **Probability density function of number of authors $\bar{a}_i$ per paper.** The vertical line falls at ten authors, corresponding roughly to the point were the distribution deviates from the power law fitting line. Only the last 3% of the most authored papers are not taken into account.

[S1] Martin, T., Ball, B., Karrer, B. & Newman, M. E. J. Coauthorship and citation patterns in the physical review. *Phys. Rev. E* **88**, 012814 (2013). URL http://link.aps.org/doi/10.1103/PhysRevE.88.012814.

[S2] http://nces.ed.gov/.