CERN-TH-2018-066

Biblioranking fundamental physics (updated to 2021/1/1)

Alessandro Strumia[a], **Riccardo Torre**[b,c]

[a] *Dipartimento di Fisica dell'Università di Pisa, Italy*
[b] *CERN, Theory Division, Geneva, Switzerland*
[c] *INFN, sezione di Genova, Italy*

Abstract

Counting of number of papers, of citations and the h-index are the simplest bibliometric indices of the impact of research. We discuss some improvements. First, we replace citations with *individual citations*, fractionally shared among co-authors, to take into account that different papers and different fields have largely different average number of co-authors and of references. Next, we improve on citation counting applying the PageRank algorithm to citations among papers. Being time-ordered, this reduces to a weighted counting of citation descendants that we call *PaperRank*. We compute a related *AuthorRank* applying the PageRank algorithm to citations among authors. These metrics quantify the impact of an author or paper taking into account the impact of those authors that cite it. Finally, we show how self- and circular- citations can be eliminated by defining a closed market of *Citation-coins*. We apply these metrics to the InSpire database that covers fundamental physics, presenting results for papers, authors, journals, institutes, towns, countries, and continents, for all-time and in recent time periods.

Contents

1 Introduction **2**

2 Ranking papers **7**
2.1 PaperRank . . . 7
2.2 PaperRank of papers: results . . . 8
2.3 PaperRank as the number of citations-of-citations . . . 10
2.4 Top-referred (recent) papers . . . 11
2.5 Paper metrics: correlations . . . 13

3 Ranking authors **14**
3.1 Sharing among co-authors: fractional counting . . . 16
3.2 Fractional counting and collaborations . . . 16
3.3 PaperRank of authors: results . . . 17
3.4 Author Rank . . . 19
3.5 Removing self-citations and citation 'cartels': the Citation-coin . . . 20
3.6 Author metrics: correlations . . . 22

4 Rankings groups **24**
4.1 Ranking institutions . . . 24
4.2 Ranking towns . . . 26
4.3 Ranking countries and continents . . . 26
4.4 Ranking journals . . . 27

5 Conclusions **28**

A The InSpire and arXiv databases **34**
A.1 Main trends in the fundamental physics literature . . . 34
A.2 Details about the dataset . . . 36

1 Introduction

Bibliometrics can be a useful tool for evaluating research: it provides simple, quick, first objective measures of the impact of papers and authors and is increasingly being considered a useful (although incomplete) evaluation criterion in postdoc/faculty recruitments, fundings, and grant awards (Henneken and Kurtz, 2017, Kurtz, 2017, Kurtz and Henneken, 2017). In the fundamental physics community that we consider in this paper, the most common measures of the impact such as counting of number of papers, citations and Hirsch's h-index (Hirsch, 2005), are inflating (Sinatra, Deville, Szell, Wang, and Barabási, 2015), making it harder to identify the real impact of research, especially for the most recent literature. The more papers one writes and the more citations these papers get, the bigger bibliometric estimators become: it does not matter if these citations close in certain loops and/or remain confined in sub-fields,

or whether the paper has been written by a single author or in collaboration with thousands of people.

We introduce new metrics and compare them with the existing ones, showing how they address the issues mentioned above. We apply them to the INSPIRE [1] bibliographic database (Holtkamp, Mele, Simko, and Smith, 2010, Ivanov and Raae, 2010, Klem and Iwaszkiewicz, 2011, Martin Montull, 2011),[2] that covers fundamental physics literature after $\approx$ 1970.[3] The metrics, both the usual ones and the new ones that we introduce, can measure the impact of papers, $p, p', \ldots$, of authors, $A, A', \ldots$, and of groups. They are defined as follows:

1. **Number of papers**
   The most naive metric consists in counting the number of papers $N_A^{\text{pap}} = \sum_{p \in A} 1$ written by a given author A. This metric rewards the most prolific authors.

2. **Number of citations**
   The most used metric consists in counting the number of citations N_p^{cit} received by a paper p. An author A is then evaluated summing the number of citations N_A^{cit} received by its papers. In formulæ:

$$N_p^{\text{cit}} = \sum_{p' \to p} 1 \,, \qquad N_A^{\text{cit}} = \sum_{p \in A} N_p^{\text{cit}} \,, \tag{1}$$

   where the first sum runs over all papers p' that cite p, and the second sum over all papers p of author A.

3. **h-index**
   The h-index is defined as the maximum h such that h papers have at least h citations each. In formulæ, assuming that all papers of author A are sorted in decreasing order of number of citations $N_p^{\text{cit}} \geq N_{p+1}^{\text{cit}}$, it is given by

$$h_A = \max \left\{ p \,|\, p \leq N_p^{\text{cit}} \right\} \,. \tag{2}$$

   This is proportional to $\sqrt{N_A^{\text{cit}}}$, times a factor that penalises authors that write a small number of highly cited papers (Hirsch, 2005).

[1] https://inspirehep.net.

[2] Other notable digital libraries and databases of research literature in various fields are, for instance, ADS - The SAO/NASA Astrophysics Data System (http://www.adsabs.harvard.edu), CDS - CERN Document Server (https://cds.cern.ch), arXiv.org (https://arxiv.org), Google Scholar (https://scholar.google.it), Microsoft Academic (https://academic.microsoft.com), DBLP - Computer Science Bibliography (https://dblp.uni-trier.de), ACM Digital Library (https://dl.acm.org), PubMed (https://www.ncbi.nlm.nih.gov/pubmed), MathSciNet - Mathematical Reviews (https://mathscinet.ams.org), CiteSeerX (https://citeseerx.ist.psu.edu), Semantic Scholar (https://www.semanticscholar.org), RePEc - Research Papers in Economics (http://repec.org), IEEE Xplore Digital Library (https://ieeexplore.ieee.org), and zbMATH the first resource for mathematics (https://zbmath.org).

[3] The most relevant literature before $\approx$ 1970 has been added and is still being added on a request base to INSPIRE.

As we will see, the average number of authors per paper and of references per paper increased, in the last 20 years, by one to a few per-cent a year, and is significantly different in different communities. Following basic common-sense, we propose an improved metric that renormalises away such factors, and that cannot be artificially inflated adding more references and/or more co-authors.

4. **Number of individual citations**
   A citation from paper p' to paper p is weighted as the inverse of the number of references $N_{p'}^{\text{ref}}$ of paper p'. Furthermore, the citations received by a paper p are equally shared among its N_p^{aut} authors.[4] In formulæ:

$$N_p^{\text{icit}} = \sum_{p' \to p} \frac{1}{N_{p'}^{\text{ref}}}\,, \qquad N_A^{\text{icit}} = \sum_{p \in A} \frac{N_p^{\text{icit}}}{N_p^{\text{aut}}}\,. \tag{3}$$

The definition of individual quantities is not new to the scientometrics literature and has been extensively studied as an application of the so-called *fractional counting*. Fractional counting has been considered both in the context of metrics and rankings (Aksnes, Schneider, and Gunnarsson, 2012, Bouyssou and Marchant, 2016, Carbone, 2011, Egghe, 2008, Hooydonk, 1997, Leydesdorff and Bornmann, 2010, 2011, Leydesdorff and Opthof, 2010, Leydesdorff and Shin, 2011, Rousseau, 2014), and in the context of constructing research networks (Leydesdorff and Park, 2017, Perianes-Rodríguez, Waltman, and van Eck, 2016).

Furthermore, as well known in the literature, the division by $N_{p'}^{\text{ref}}$ factors out the different publication intensity in the various fields, without the need of a field classification system. Indeed, a paper has $N_p^{\text{icit}} = 1$ if it receives the mean number of citations in its field.

All the above metrics are defined "locally", i.e. they can be computed for a paper/author without knowing anything about all other papers but the ones that cite it. Therefore, they all potentially suffer from the problem that even small sub-communities can inflate their indicators. To overcome this problem one needs to define global measures, i.e. measures that know about the whole community. The simplest such global measure of impact is given by the *PageRank* algorithm, introduced in 1996 by Larry Page and Sergey Brin,[5] the founders of Google (Brin, Larry Page, and Winograd, 1999, Brin and Page, 1998).[6] For a pedagogical introduction to the PageRank see Rajaraman and Ullman (2009) and Tanase and Radu (2009). Applications of the PageRank algorithm to citations network have already been considered, for instance, by Chen, Xie, Maslov, and Redner (2007), Ding, Yan, Frazho, and Caverlee (2009), Ma, Guan, and Zhao (2008), and Zhou, Orshanskiy, Zha, and Giles (2007). More advanced ranking algorithms, based on integrated bibliographic information have also been proposed by Bini, del Corso, and Romani (2010), Bini, del Corso, and Romani (2008), and del Corso and Romani (2009).

[4] We assume that authors contributed equally because in fundamental physics authors are usually listed alphabetically, with no information about who contributed more. The factor $1/N_p^{\text{aut}}$, that is the one dictated by conservation laws, will be further motivated in Section 3.2.

5. **PaperRank**
   The PaperRank R_p of paper p and the PaperRank R_A of author A are defined as

$$\boxed{R_p = \sum_{p' \to p} \frac{R_{p'}}{N^{\rm ref}_{p'}}\,, \qquad R_A = \sum_{p \in A} \frac{R_p}{N^{\rm aut}_p}}\,. \tag{4}$$

   Namely, citations from papers p' are weighted proportionally to their ranks $R_{p'}$, that get thereby determined trough a system of linear equations. The PaperRank provides a metric which cannot be easily artificially inflated, because it is the bibliometric estimator of a physical quantity: how many times each paper is read.

As we will see, the PaperRank singles out notable old papers which often do not have many citations. However, given that citations are time-ordered (newer papers cite older ones), the rank reduces to a weighted sum over citation descendants (a combination of "citations of citations") which needs about 10-20 years to become a better indicator than the number of individual citations. In order to use information from the past, we define an alternative AuthorRank based on citations among authors.

6. **AuthorRank**
   We define the citation matrix which counts all individual citations from author A' to A

$$N^{\rm icit}_{A' \to A} = \sum_{p_{A'} \to p_A} \frac{1}{N^{\rm aut}_{p_A} N^{\rm aut}_{p_{A'}}} \frac{1}{N^{\rm ref}_{p_{A'}}}\,, \tag{5}$$

   where the sum runs over all papers $p_{A'}$ of author A' that cite papers p_A of A. We then define the AuthorRank $\mathscr{R}_A$ as

$$\boxed{\mathscr{R}_A = \sum_{A'} \mathscr{R}_{A'} C_{A' \to A}\,, \qquad C_{A' \to A} = \frac{N^{\rm icit}_{A' \to A}}{\sum_{A''} N^{\rm icit}_{A' \to A''}}}\,, \tag{6}$$

   namely, as the principal eigenvector of the right stochastic matrix $C_{A' \to A}$, which (thanks to the normalization of each row provided by the sum in the denominator) tells the percentage of individual citations to A among all individual citations of A'. The AuthorRank gives more weight to citations coming from highly cited authors. We also use the AuthorRank of authors to define an improved ranking of papers, that we call AuthorRank of papers, as

$$\boxed{\mathscr{R}_p = \sum_{p' \to p} \sum_{A \in p'} \frac{\mathscr{R}_A}{N^{\rm aut}_{p'} N^{\rm ref}_{p'}}}\,. \tag{7}$$

[5] A similar idea, applied to bibliometrics, have been proposed long before the advent of the PageRank by Pinski and Narin (1976).

[6] Even if the word "Page" in PageRank may seem to refer to webpages, the name of the algorithm originates from the name of one of its inventors, Larry Page.

Ideas similar to our AuthorRank have already been considered in the literature (Radicchi, Fortunato, Markines, and Vespignani, 2009, West, Jensen, Dandrea, Gordon, and Bergstrom, 2013). Radicchi et al. (2009) applied an algorithm similar to our AuthorRank to the Physical Review publication archive up to 2006. The implementation, called Science Author Rank Algorithm (SARA) is different than ours in few respects: first, it is based on slices of the full database, given by multiple graphs with equal number of citations, while our AuthorRank includes information from a single graph constructed with weighted citations among all authors after a given time (or for all times). Second, the SARA authors consider about half the number of papers we consider and around one third of the citations we consider, all coming from the single publisher Physical Review. While the analysis of Radicchi et al. (2009) is pioneering in this direction, we believe that focusing on a single publisher introduces bias, since it does not cover the whole scientific production of scientists: only a fraction of each author's papers are published on a Physical Review journal. We avoid this bias by considering the InSpire public database including the scientific research in Fundamental Physics consisting of preprints and articles published on about $2 \cdot 10^3$ journals. This should remove some of the bias induced by considering a single publisher.

Third and most important difference between SARA and the AuthorRank is the contribution of citations among papers to the weight of the links in the author-level graph: in the case of SARA each paper contributes to the author-level graph with a weight $1/(m \cdot n)$, with m and n the number of authors in the citing and cited paper. In our case instead this weight is $1/N^{\rm icit}_{A' \to A}$, with this quantity defined in Eq. (5). As can be seen by this equation, this also contains a $1/N^{\rm ref}_{p_{A'}}$. In simpler words, the SARA author-level graph is determined using citations among papers, in our notation $N^{\rm cit}_p$, while our author-level graph is determined using individual citations among papers, in our notation $N^{\rm icit}_p$, as defined in Eq. (3). We find the latter more indicative, as we find individual citations of papers, that are the building blocks of the paper-level graph, more indicative than traditional citations. Being cited among many others is typically less relevant than being cited with a few other references.

West et al. (2013) applied a similar algorithm to the Social Science Research Network Community, calling it Author-Level Eigenfactor Metric. In this case some bias is introduced by removing all self-citations from the graph, corresponding to the diagonal of the correlation matrix $N^{\rm icit}_{A' \to A}$. As we discuss in Section 3.5, removing self-citations is an arbitrary procedure, also admitting different implementations. The problem of making the effect of self citations less relevant can be addressed in two different ways: by tuning the teleport probability parameter in the PageRank algorithm, or by considering a more general quantity, like the Citation-coin introduced below.

In this paper we stress that, while the AuthorRank is an interesting metric by itself, it mainly identifies already well known authors, like Feynman and others (Radicchi et al., 2009), and Jensen and others (West et al., 2013). We try to improve in extracting information from the AuthorRank and use it to identify recent papers cited by highly ranked (with AuthorRank) physicists. This allows us to define an AuthorRank for papers, that is a good candidate as an early alert for potentially important papers.

As we stated before, removing self-citations carries some level of arbitrariness. One could aim at resolving a better defined problem, consisting of removing all citation 'cartels', defined

as loops of citations among 3 authors, among 4 authors, etc. The mathematical problem of removing all circular citations has a simple solution, inspired by economy: money.[7] You don't get richer by giving money to yourself or by circulating money with friends. This leads us to the definition of an additional metric:

7. **Citation-coin**
   Author A 'owes' the number $\not\!\!\!C_A$ of individual citations received minus the number of individual citations given:[8]

$$\not\!\!\!C_A = \sum_{A'} (N^{\text{icit}}_{A'\to A} - N^{\text{icit}}_{A\to A'}) = N^{\text{icit}}_A - \sum_{p\in A} \frac{1}{N^{\text{aut}}_p} \,. \tag{8}$$

   This metric penalises authors who write many small papers which receive few citations from others.

An important property of all above metrics is that they can be computed in practice. Our metrics are intensive: so they can be used to rank groups, such as journals, institutes, countries, etc. by simply summing over their members. Furthermore they can be restricted to any specific time period, e.g. after year 2000.

The paper is structured as follows. In Section 2 we introduce the PaperRank R, discuss its features and properties, and rank all papers in InSpire. In Section 3 we introduce the number of individual citations N_{icit}, the AuthorRank $\mathscr{R}$ and the Citation-coin $\not\!\!\!C$, discussing their features and properties, and rank all authors in InSpire. In Section 4 we apply these measures to rank groups: institutions, towns, countries, continents, and journals. Conclusions are presented in Section 5. In Appendix A we describe the InSpire and arXiv databases and their main features and trends, together with technical details.

Several of our results with complete tables are available at the PhysRank webpage[9].

2 Ranking papers

2.1 PaperRank

Given a citation network of N_{pap} papers, the PaperRank R_p of each paper p is defined by

$$R_p = \wp \sum_{p'\to p} \frac{R_{p'}}{N^{\text{ref}}_{p'}} + \alpha(1-\wp)\,, \tag{9}$$

where $N^{\text{ref}}_{p'}$ is the total number of references of each paper p' that cites p. EquationEq. (9) is linear, so its solution is unique, and can be computed efficiently iteratively (Rajaraman and Ullman, 2009).

[7] In mathematical language this consists in removing all closed loops from the authors' graph, making it acyclic.

[8] A slight improvement will be added later, to avoid border effects due to intrinsic finite nature of the database.

[9] http://rtorre.web.cern.ch/rtorre/PhysRank.

We elaborate on its meaning. Eq.Eq. (9) contains two arbitrary constants α and $\wp$. The constant α just fixes the overall normalization $\sum_p R_p = R_{\rm tot}$. When applied to internet, R_p describes the probability that site p is visited and it is convenient to normalize it to one. We choose $R_{\rm tot}$ equal to the total number of citations $R_{\rm tot} = \sum_p N_p^{\rm cit}$, in order to allow for an easier comparison between the number of citations received by a paper and its PaperRank R_p. With this normalisation R_p grows with time, as newer papers appear.

Viewing the rank as the probability that a paper is read, the parameter $\wp$ splits it into two contributions: the first term is the probability that a reader reaches a paper by following a reference to it; the second term, equal for all papers, simulates readers that randomly browse the literature.

- In the limit $\wp = 0$ the first contribution vanishes, and all papers have a common rank. At first order in small $\wp \ll 1$, R_p starts discriminating the papers p:

$$R_p(\wp) \overset{\wp \ll 1}{\simeq} \frac{R_{\rm tot}}{N_{\rm pap}} \frac{1 + \wp N_p^{\rm icit}}{1 + \wp} \,, \tag{10}$$

  where $N_p^{\rm icit}$ is the number of individual citations received by p defined in Eq. (3), which obeys $\sum_p N_p^{\rm icit} = N_{\rm pap}$.

- In the limit $\wp = 1$ the second contribution in Eq. (9) vanishes, and R_p only depends on the structure of the network, provided that no closed sub-networks and dead-ends exist (Rajaraman and Ullman, 2009). Recursive computations of R_p become slower as $\wp \to 1$.

Data about downloads of scientific articles would allow to extract the value of $\wp$ that better fits the observed reading rate; however such data are not available in fundamental physics.[10] We use a large $\wp = 0.99$, such that the first contribution in Eq. (9) dominates for all relevant authors.

2.2 PaperRank of papers: results

We compute the PaperRank by constructing a graph (and its transition matrix) having all papers as nodes and citations as links. We consider the full InSpire database, as detailed in Appendix A.2. Generally, a few hundred iterations of Eq. (9) are necessary for a percent level convergence. The computation takes a few minutes on a laptop computer.

Table 1 (Table 2) shows the top-cited (top-ranked) papers in the InSpire database. Top-ranked papers correspond to the papers with top PaperRank and tend to be old famous ones, even with a relatively small number of citations. Top-cited papers, ranked with the usual counting of the number of citations, tend to be modern, in the view of the inflation in the rate of citations. The same effect was observed by Chen et al. (2007), who applied the PageRank algorithm to the sub-set of papers published on Physical Review.

[10] arXiv.org does not make public the number of downloads, to avoid the conversion of this information into a relevant metric, and its consequent fate determined by Goodhart's law.

	Title	1st author	$N_{\rm aut}$	date	$N_{\rm cit}$	R_p	$\mathscr{R}_p$
1	*The Large N limit of superconformal fiel*	J.M.Maldacena	1	1998	16317	6725	22034
2	*GEANT4–a simulation toolkit*	James.R.Allison	127	2003	13488	6790	2170
3	*Measurements of Ω and Λ f*	S.Perlmutter	32	1999	12983	4124	7543
4	*Observational evidence from supernovae f*	A.G.Riess	20	1998	12965	4267	7376
5	*A Model of Leptons*	Steven.Weinberg	1	1968	12955	44908	41591
6	*PYTHIA 6.4 Physics and Manual*	T.Sjostrand	3	2006	11823	3439	3798
7	*Observation of a new particle in the sea*	ATLAS	2932	2012	11666	1652	4074
8	*Observation of a New Boson at a Mass of*	CMS	2897	2012	11404	1638	4014
9	*CP Violation in the Renormalizable Theor*	M.Kobayashi	2	1973	10695	10903	20359
10	*Anti-de Sitter space and holography*	E.Witten	1	1998	10513	4471	13998

Table 1: *Top-cited (highest number of citations) papers in the* INSPIRE *database.*

	Title	1st author	$N_{\rm aut}$	date	$N_{\rm cit}$	R_p	$\mathscr{R}_p$
1	*A Model of Leptons*	Steven.Weinberg	1	1968	12955	44908	41591
2	*Conservation of Isotopic Spin and Isotop*	C.N.Yang	2	1954	2852	41770	14876
3	*Theory of Fermi interaction*	R.P.Feynman	2	1958	1748	39446	10778
4	*Remarks on the Dirac theory of the posit*	W.Heisenberg	1	1934	122	38682	1223
5	*On the Stopping of fast particles and on*	H.A.Bethe	2	1934	724	31585	1009
6	*The S matrix in quantum electrodynamics*	F.J.Dyson	1	1949	771	31296	5044
7	*Symmetries of baryons and mesons*	M.Gell.Mann	1	1962	1666	29947	8576
8	*Field Theories with Superconductor Solut*	J.Goldstone	1	1961	2078	29244	5773
9	*A Theory of the Fundamental Interactions*	J.S.Schwinger	1	1957	679	28331	11962
10	*Space - time approach to quantum electro*	R.P.Feynman	1	1949	830	26963	3550

Table 2: *Top-ranked (highest PaperRank) papers in the* INSPIRE *database.*

	Title	1st author	$N_{\rm aut}$	date	$N_{\rm cit}$	R_p	$\mathscr{R}_p$
1	*A Model of Leptons*	Steven.Weinberg	1	1968	12955	44908	41591
2	*Particle Creation by Black Holes*	S.W.Hawking	1	1974	8665	9142	28472
3	*A Planar Diagram Theory for Strong Inter*	G.tHooft	1	1974	4989	8238	26994
4	*Unity of All Elementary Particle Forces*	H.M.Georgi	2	1974	5090	10397	25424
5	*Confinement of Quarks*	K.G.Wilson	1	1974	5376	23570	23192
6	*Weak Interactions with Lepton-Hadron Sym*	S.L.Glashow	3	1970	6214	19199	22997
7	*Pseudoparticle Solutions of the Yang-Mil*	A.A.Belavin	4	1975	2887	13352	22315
8	*The Large N limit of superconformal fiel*	J.M.Maldacena	1	1998	16317	6725	22034
9	*CP Violation in the Renormalizable Theor*	M.Kobayashi	2	1973	10695	10903	20359
10	*Symmetry Breaking Through Bell-Jackiw An*	G.tHooft	1	1976	3769	8220	20081

Table 3: *Top-referred (highest AuthorRank) papers in the* INSPIRE *database.*

The difference between the two rankings is partly due to the fact that PaperRank penalises recent papers. Papers tend to accumulate citations for about 10-20 years, while the rank continues growing with time, and is highly suppressed for younger papers.

This also means that the PaperRank defined in Eq. (9) needs 10-20 years before providing a better metrics than the number of citations. This is proven in the next section, where we show that, for a time-ordered network (such as the network of citations), the PaperRank reduces to the number of citations-of-citations.

2.3 PaperRank as the number of citations-of-citations

Internet allows for reciprocal links among pages, and the PageRank captures in a simple way the self-interacting system. Citations among scientific papers are instead time-ordered, forming an acyclic network. In the limit where citations of older papers to newer papers are ignored,[11] no loops are possible within the network, and the implicit definition of the rank R_p of Eq. (9) can be converted into the following explicit expression[12]

$$R_p \propto \sum_{g=0}^{\infty} \wp^g \sum_{p_g \to \cdots \to p} \frac{1}{N^{\rm ref}_{p_g}} \cdots \frac{1}{N^{\rm ref}_{p_1}} \,. \tag{11}$$

Basically, R_p counts the number of citations-of-citations up to generation g. In the above expression, the term with

- $g = 0$ contributes as unity, and accounts for the constant term in Eq. (9), which is negligible for papers that receive citations from others;
- $g = 1$ contributes with the number of individual citations $N^{\rm icit}_p$ as in Eq. (10): the sum runs over 'first generation' papers p_1 that cite the paper p;
- $g = 2$ corresponds to 'second generation' papers p_2 that cite the papers p_1 that cite p;
- $g = 3$ corresponds to 'third generation' papers p_3 that cite the papers p_2 that cite the papers p_1 that cite p;
- for generic g the sum runs over all papers p_g that cite paper p in g steps.

In other terms, for any given paper, we refer to papers that cite it as "first generation", and define as "second generation" those papers that cite at least a first generation paper, and so on. A paper q can appear multiple times in different generations g, corresponding to all possible citations paths from q to p. Eq.Eq. (11) shows that $\wp < 1$ gives a cut-off on the number of generations that one wants to consider, and that $R_p(\wp)$ grows with $\wp$, and with time.

[11] We enforced time-ordering within the citations, deleting from the InSpire database a small number of 'a-causal' citations, where older papers cite newer papers (see Appendix A.2 for details). Since older papers tend to accumulate large ranks, a-causal citations can artificially inflate the rank of a few recent papers.

[12] This can be proven by substituting Eq. (11) into Eq. (9). A physicist can view in Eq. (11) a path-integral within the network.

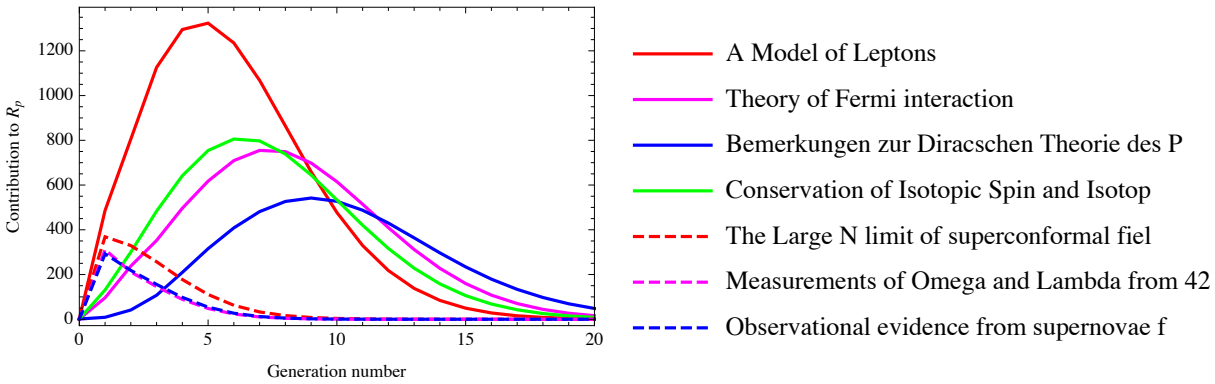

Figure 1: *Contributions of each generation to the citation chain of some notable papers.*

- At one extremum, $\wp \to 0$, the rank R_p reduces to the "number of children" N_p^{icit}, without checking if they are successful. Papers on hot topics can fast accumulate many citations, even if later the hot topic becomes a dead topic. Too recent papers are penalised.

- At the other extremum, $\wp \to 1$, the rank R_p becomes the Adamo number: it counts descendants. Seminal papers that open new successful fields are rated highly, and their rank continues to grow. Too recent papers are highly penalised.

The PaperRank in Eq. (11) splits papers into two qualitatively different categories:

- Sub-critical papers that get some attention and get forgotten: this happens when, after a long time (tens of years), the sum over g remains dominated by the first generation.

- Super-critical papers, that make history. If the citation rate is high enough, it can sustain a 'chain reaction', such that late generations keep contributing significantly to the sum over generations. At the same time, the original paper gets summarized in books and ceases to be cited directly.

Figure 1 shows, for a few notable papers, how much different generations g contribute to the sum in Eq. (11). We see that about 10 generations contribute significantly for old top-ranked papers, while the 1st generation provides the dominant contribution to recent top-cited papers.[13]

2.4 Top-referred (recent) papers

As explained in the previous section, the PaperRank can single out some notable papers with few citations, provided that they are old. However, when applied to recent papers (less than 10-

[13]We computed $R_p(\wp)$ analytically as function of $\wp$ for all papers p using Eq. (11) with the following 'pruning' algorithm. To start, one finds all papers with no citations and eliminates them from the database, after assigning their contributions to the R_p of their references. The process is iterated. About 1000 iterations are needed to prune all the InSpire citation tree to nothing, obtaining $R_p(\wp)$ as a power series in $\wp$.

Year	Title	1st author	$N_{\rm aut}$	$N_{\rm cit}$	R_p	$\mathscr{R}_p$
1880	*On the Relative Motion of the Earth and the*	A.A.Michelson	2	211	219	127
1890	*Cathode rays*	J.J.Thomson	1	239	25	33
1900	*On the electrodynamics of moving bodies*	Albert.Einstein	1	499	754	1272
1910	*The Foundation of the General Theory of Rel*	Albert.Einstein	1	1104	1415	973
1910	*Approximative Integration of the Field Equa*	Albert.Einstein	1	331	1552	6501
1910	*Einstein's theory of gravitation and its as*	de Sitter	1	219	3474	182
1920	*About the Pauli exclusion principle*	E.P.Wigner	2	383	23416	2222
1920	*Quantum Theory and Five-Dimensional Theory*	O.Klein	1	2481	3233	5741
1930	*Quantised singularities in the electromagne*	P.A.M.Dirac	1	2269	20421	12586
1930	*Remarks on the Dirac theory of the positron*	W.Heisenberg	1	122	38682	1223
1940	*The Theory of magnetic poles*	P.A.M.Dirac	1	1071	6126	6821
1940	*The S matrix in quantum electrodynamics*	F.J.Dyson	1	771	31296	5044
1940	*Forms of Relativistic Dynamics*	P.A.M.Dirac	1	1828	1643	4095
1950	*On gauge invariance and vacuum polarization*	J.S.Schwinger	1	5160	10263	11589
1950	*Conservation of Isotopic Spin and Isotopic*	C.N.Yang	2	2852	41770	14876
1960	*A Model of Leptons*	Steven.Weinberg	1	12955	44908	41591
1970	*CP Violation in the Renormalizable Theory o*	M.Kobayashi	2	10695	10903	20359
1970	*Confinement of Quarks*	K.G.Wilson	1	5376	23570	23192
1970	*Particle Creation by Black Holes*	S.W.Hawking	1	8665	9142	28472
1980	*The Inflationary Universe: A Possible Solut*	A.H.Guth	1	8171	9250	15347
1990	*The Large N limit of superconformal field t*	J.M.Maldacena	1	16317	6725	22034
2000	*GEANT4–a simulation toolkit*	GEANT	127	13488	6790	2170
2000	*First year Wilkinson Microwave Anisotropy P*	WMAP	17	9040	2990	6507
2010	*Observation of a new particle in the search*	ATLAS	2932	11666	1652	4074

Table 4: *The top-cited, top-ranked, top-referred paper written in each decennium among those listed in* InSpire *(which is highly incomplete before 1960).*

20 years old), the PaperRank becomes highly correlated to the number of (individual) citations, and therefore cannot perform better.

We thereby propose an early-alert indicator that recovers some information from the past. First, we compute a rank among authors using all-time data. In particular, we adopt the AuthorRank $\mathscr{R}_A$ anticipated in Eq. (6) and better discussed in the next section. Next, we use such rank to weight citations to papers, as in Eq. (7). This means that we give more weight to citations from authors with higher $\mathscr{R}_A$, implementing a sort of representative democracy. We dub papers with top AuthorRank of papers $\mathscr{R}_p$ as 'top-referred papers'. Table 3 shows the all-time list.

Table 4 shows the top-cited, top-ranked and top-referred papers published within each decennium, based on all subsequent citations. For recent papers, top-cited and top-ranked tend to be dominated by manuals of useful computer codes and by reviews.

We finally use $\mathscr{R}_p$ to find top-referred recent papers: Table 5 shows the top-referred papers published after year 2010 and with less than 10 authors (because our goal is to identify notable recent papers less known than discoveries made by big experimental collaborations) and more than 100 citations. Furthermore, we here removed self-citations, to avoid the list to be

	Title	1st author	$N_{\rm aut}$	date	$N_{\rm cit}$	R_p	$\mathscr{R}_p$
1	*Black Holes: Complementarity or Firewall*	A.Almheiri	4	2012	1129	230	4712
2	*New Symmetries of QED*	D.Kapec	3	2015	137	36	1725
3	*Conformal symmetry and its breaking in t*	J.M.Maldacena	3	2016	504	84	1092
4	*On the Origin of Gravity and the Laws of*	E.P.Verlinde	1	2010	880	196	1082
5	*Remarks on the Sachdev-Ye-Kitaev model*	J.M.Maldacena	2	2016	860	173	1059
6	*Topological insulators and superconducto*	X.L.Qi	2	2011	1592	534	1026
7	*Black holes and the butterfly effect*	S.H.Shenker	2	2013	733	198	939
8	*Topological Insulators*	M.Zahid.Hasan	2	2010	1892	788	888
9	*Shapiro Delay Measurement of A Two Solar*	P.Demorest	5	2011	2411	396	846
10	*Investigating the near-criticality of th*	D.Buttazzo	7	2014	1086	127	845

Table 5: *Top-referred papers written after 2010 with less than 10 authors and more than 100 citations.*

	$N_p^{\rm cit}$	$N_p^{\rm icit}$	R_p	$\mathscr{R}_p$
$N_p^{\rm cit}$		0.95	0.43	0.72
$N_p^{\rm icit}$			0.49	0.78
R_p				0.63
$\mathscr{R}_p$				

	$N_A^{\rm pap}$	$N_A^{\rm cit}$	h_A^2	$N_A^{\rm ipap}$	$N_A^{\rm icit}$	R_A	$\mathscr{R}_A$	$\mathscr{C}\!\!\!/_A$
$N_A^{\rm pap}$		0.95	0.95	0.19	0.08	0.04	0.06	0.03
$N_A^{\rm cit}$			0.98	0.09	0.08	0.04	0.06	0.07
h_A^2				0.08	0.06	0.02	0.05	0.05
$N_A^{\rm ipap}$					0.64	0.36	0.46	0.25
$N_A^{\rm icit}$						0.72	0.83	0.88
R_A							0.85	0.71
$\mathscr{R}_A$								0.82
$\mathscr{C}\!\!\!/_A$								

Figure 2: *Left(right) table: correlations between indices of papers(authors). For papers the correlation is considered separately for the whole* INSPIRE *and for the eight main arXiv categories.*

dominated by notable authors citing themselves.

2.5 Paper metrics: correlations

The left panel of Fig. 2 shows the correlations among the traditional counting of citations $N_p^{\rm cit}$, the number of individual citations $N_p^{\rm icit}$, the PaperRank R_p and the AuthorRank of papers $\mathscr{R}_p$, in the whole INSPIRE database. The number of individual citations is highly correlated with the number of citations. Indeed, for papers, individual citations are just citations divided by the number of references of the citing papers so that uncorrelation is proportional to the variance around the average of the number of references per paper. The PaperRank and the AuthorRank are less correlated to the number of (individual) citations and to each other and represent fairly independent indices for ranking papers.

Number of papers $\sum_{\text{papers}} 1$

	InSpires name	All InSpires
1	G.Eigen.1	2508
2	S.L.Wu.1	2396
3	Kazuhiko.Hara.1	2359
4	J.Brau.2	2310
5	A.Seiden.1	2293
6	R.Kass.1	2276
7	David.M.Strom.1	2270
8	A.Bodek.2	2268
9	W.T.Ford.1	2263
10	R.V.Kowalewski.1	2231

Number of citations $\sum_{\text{papers}} N_{\text{cit}}$

	InSpires name	All InSpires
1	R.V.Kowalewski.1	261176
2	G.D.Cowan.1	247656
3	Otmar.Biebel.1	236405
4	J.Huston.1	241391
5	Christoph.Grab.1	225873
6	Achim.Stahl.1	220954
7	K.Moenig.1	222876
8	S.L.Wu.1	225586
9	S.M.Spanier.1	215374
10	A.V.Gritsan.1	217487

h index $\sum_{N_{\text{cit}} \geq h} 1$

	InSpire name	All InSpires
1	Kazuhiko.Hara.1	203
2	J.Huston.1	200
3	H.H.Williams.1	198
4	S.L.Wu.1	197
5	A.G.Clark.1	196
6	G.Eigen.1	196
7	P.K.Sinervo.1	195
8	M.J.Shochet.1	195
9	J.Proudfoot.1	195
10	S.M.Errede.1	195

Average citations $\langle N_{\text{cit}} \rangle$

	InSpire name	All InSpire
1	Y.Oohata.1	13488
2	N.Eiden.1 et al.	7673
3	J.L.Chuma.1 et al.	6744
4	S.Chowdhury.1 et al.	6389
5	P.Schaffner.1	6104
6	S.B.Lugovsky.1	6043
7	K.S.Lugovsky.1	5844
8	C.Roumenin.1 et al.	5746
9	H.Yusupov.2	5732
10	V.S.Lugovsky.1	5195

Table 6: *Authors listed according to traditional biblio-metric indices: total number of papers (top left), of citations (top right), h-index (bottom left), average number of citations per paper (bottom right).*

3 Ranking authors

We start from the simplest and most naive metrics: in the top-left column of Table 6 we list the authors with most papers. Within the InSpire database, they are all experimentalists that participate in large collaborations with many co-authors. The extreme case are the ATLAS and CMS collaborations with $\sim 10^3$ papers and $\sim 10^3$ authors.

In the top-right column of Table 6 we show the top-cited authors: again they are experimentalists that participate in large collaborations. The citations of author A are counted in the usual way: summing the citations received by all papers that include A as author, as in Eq. (1). The bottom-left column of Table 6 shows the authors with highest h index, and the bottom-right column the authors with the highest average number of citations per paper.

Next to each author we add symbols which show if they received the Nobel (), Dirac (from ICTP and from IOP), Planck (), Sakurai (), Wolf () and Milner () prizes. Small inaccuracies are possible, as medalists have been identified from names. None of the top authors

	InSpire name	All	InSpire name	After 2000	InSpire name	After 2010
1	E.Witten.1	3703	J.M.Maldacena.1	306.1	S.D.Odintsov.1	82.9
2	Steven.Weinberg.1	2487	S.D.Odintsov.1	276.1	J.M.Maldacena.1	82.3
3	G.tHooft.1	1592	E.Witten.1	272.0	N.Kidonakis.1	77.1
4	S.W.Hawking.1	1382	T.Padmanabhan.1	254.7	E.Witten.1	71.8
5	A.M.Polyakov.1	986.2	P.Z.Skands.1	254.4	U.G.Meissner.1	63.3
6	F.A.Wilczek.1	902.6	T.Sjostrand.1	252.2	A.Strominger.1	62.2
7	J.M.Maldacena.1	900.8	S.Nojiri.1	246.6	C.de.Rham.1	62.1
8	R.W.Jackiw.1	888.3	S.Mrenna.1	219.2	S.Tsujikawa.1	60.4
9	J.S.Schwinger.1	854.4	G.P.Salam.1	217.1	S.Nojiri.1	59.5
10	A.D.Linde.1	833.4	V.Springel.1	216.5	S.Capozziello.1	57.2
11	T.Sjostrand.1	829.7	D.T.Son.1	212.5	D.Stanford.1	57.2
12	L.Susskind.1	799.6	Ashoke.Sen.1	208.6	P.Z.Skands.1	57.1
13	S.L.Glashow.1	788.0	C.Vafa.1	186.7	L.Susskind.1	55.3
14	H.M.Georgi.1	775.0	M.Cacciari.1	180.7	S.Sachdev.1	50.0
15	N.Seiberg.1	719.0	U.G.Meissner.1	180.6	M.Czakon.1	49.4
16	P.A.M.Dirac.1	704.1	P.Nason.1	176.7	G.P.Salam.1	48.2
17	Sidney.R.Coleman.1	675.3	Ernest.Ma.1	170.2	R.Venugopalan.1	47.7
18	David.J.Gross.1	665.9	D.E.Kharzeev.1	160.3	M.Luscher.1	47.7
19	C.Vafa.1	643.0	S.Tsujikawa.1	158.9	R.E.Kallosh.1	47.1
20	S.J.Brodsky.1	626.4	A.Strominger.1	156.3	D.W.Hooper.1	46.4
21	J.R.Ellis.1	617.5	Martin.Bojowald.1	152.6	R.B.Mann.1	45.6
22	K.G.Wilson.1	613.5	V.A.Kostelecky.1	150.7	S.Hod.1	45.4
23	Abdus.Salam.1	607.3	S.S.Gubser.1	148.4	F.Maltoni.1	45.4
24	M.Luscher.1	589.9	S.Capozziello.1	145.4	R.C.Myers.1	45.2
25	J.D.Bjorken.1	586.6	J.Polchinski.1	144.7	A.D.Linde.1	44.9
26	R.L.Jaffe.1	579.0	Alan.D.Martin.1	144.3	A.De.Felice.1	43.2
27	A.Strominger.1	569.5	G.Amelino.Camelia.1	143.7	X.L.Qi.1	43.1
28	Ashoke.Sen.1	567.5	A.Loeb.1	143.1	S.F.King.1	42.9
29	G.Veneziano.1	563.2	D.W.Hooper.1	141.1	C.Bambi.1	42.5
30	C.N.Yang.1	548.7	A.Ashtekar.1	141.1	B.Schenke.1	42.3
31	J.Polchinski.1	546.2	F.Aharonian.1	139.6	T.Padmanabhan.1	42.1
32	Rabindra.N.Mohapatra.1	540.2	N.Arkani.Hamed.1	138.3	K.Hinterbichler.1	41.9
33	S.Deser.1	526.5	Wayne.Hu.1	137.1	H.T.Janka.1	41.1
34	Alexander.Vilenkin.1	525.5	A.C.Fabian.1	136.8	E.N.Saridakis.1	40.8
35	R.P.Feynman.1	524.9	Rong.Gen.Cai.1	136.5	P.Nason.1	40.2
36	L.Wolfenstein.1	524.6	G.R.Dvali.1	136.3	A.Mitov.1	40.1
37	Stephen.Louis.Adler.1	520.4	L.Susskind.1	136.1	Bing.Zhang.1	39.9
38	M.Gell.Mann.1	519.4	E.V.Shuryak.1	135.8	U.W.Heinz.1	39.3
39	B.Zumino.1	518.6	U.W.Heinz.1	134.5	Florian.R.A.Staub.1	39.3
40	John.H.Schwarz.1	518.4	F.Karsch.1	131.7	A.Strumia.1	39.2
41	M.A.Shifman.1	505.9	N.Kidonakis.1	131.5	J.Rojo.1	39.1
42	E.V.Shuryak.1	493.8	G.Soyez.1	131.1	T.Schwetz.1	39.0
43	M.B.Wise.1	485.9	A.D.Linde.1	128.9	A.J.Buras.1	39.0
44	G.W.Gibbons.1	484.8	A.A.Tseytlin.1	128.3	A.Loeb.1	38.8
45	J.D.Bekenstein.1	484.1	A.Strumia.1	124.7	G.Soyez.1	38.7
46	A.A.Tseytlin.1	483.9	M.Visser.1	123.9	E.Oset.1	38.1
47	L.N.Lipatov.1	474.9	S.D.M.White.1	122.6	V.Cardoso.1	37.8
48	H.Leutwyler.1	471.9	Nathan.J.Berkovits.1	122.2	E.P.Verlinde.1	37.2
49	T.D.Lee.1	459.0	S.Frixione.1	122.1	S.S.Ostapchenko.1	36.9
50	N.Isgur.1	456.9	J.R.Ellis.1	120.1	P.Bozek.1	36.8

Table 7: *Authors sorted according to their number of individual citations* N_A^{icit} *for the whole* InSpire *database (left), from year 2000 (middle), and from year 2010 (right).*

according to traditional metrics received any of them.

Clearly, all the indices shown in Table 6 ceased to be relevant for experimentalists in view of the large number of co-authors. This shows the need for an improved metrics that corrects for the inflation in the number of co-authors and allows at least a naive comparison of experimentalists with the rest of the community.

3.1 Sharing among co-authors: fractional counting

An improved metrics is obtained by attributing a fraction p_A of any given paper to each author A, and imposing the sum rule $\sum p_A = 1$. The fractions p_A should tell how much each author contributed to the paper. In the absence of this information, *we assume that each co-author contributed equally, so that* $p_A = 1/N_p^{\rm aut}$.[14] This is called 'fractional counting' in the bibliometric literature.

Taking into account this factor, the total number of *'individual papers'* of author A is given by $\sum_{p\in A} 1/N_p^{\rm aut}$. The same sharing among co-authors is applied to citations. The number of *'individual citations'* received by author A is defined by summing over all its papers p taking into account that citations are shared among co-authors, and weighted inversely to the number of references:

$$N_A^{\rm icit} \equiv \sum_{p\in A} \frac{N_p^{\rm icit}}{N_p^{\rm aut}} = \sum_{p\in A} \frac{1}{(\text{Number of authors of paper } p)} \sum_{p'\to p} \frac{1}{(\text{Number of references of } p')} \,. \tag{12}$$

In the same way, we share the rank R_p of each paper equally among its authors. The rank R_p of a paper approximates a physical quantity: how many times the paper is read. The rank R_A of an author inherits the same meaning: it tells the visibility of any author A, obtained by summing the visibility of its papers p as in Eq. (4), i.e.[15]

$$R_A = (\text{PaperRank of author } A) = \sum_{p\in A} \frac{(\text{PaperRank of paper } p)}{(\text{Number of authors of paper } p)} \,. \tag{13}$$

As discussed in Section 2.1 we consider $\wp = 0.99$.

3.2 Fractional counting and collaborations

We share the number of citations received by a paper among its $N_{\rm aut}$ authors as $1/N_{\rm aut}$. Some agencies adopt various less steep functions of $N_{\rm aut}$. This raises the question: how does the total number of citations received scale, on average, with the number of co-authors?[16] Figure 3 (described in more details in the caption) provides the answer: the total number of citations

[14] Weighting authors proportionally to their AuthorRank gives a non-linear system of equations, with singular solutions where a few notable authors collect all the weight of large collaborations. Restricting to single-author papers would uniquely identify authors' contributions, but at the price of discarding most literature.

[15] For different methods see Ding et al. (2009) and Zhou et al. (2007).

[16] We thank Paolo Rossi for raising this issue.

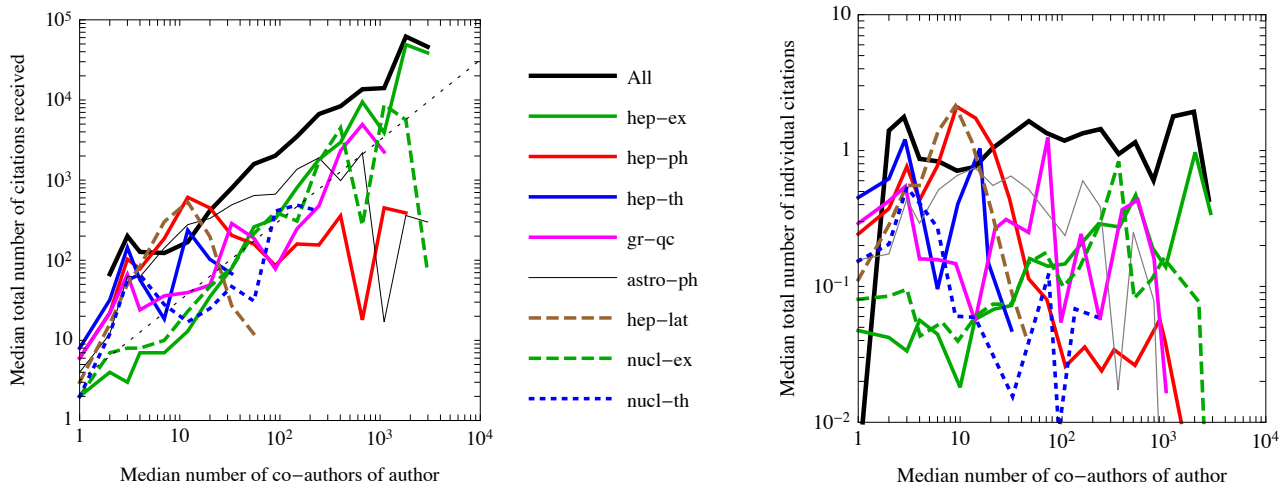

Figure 3: **Left**: *for each author in* InSpire *we compute its mean number of co-authors (per paper, including the author itself) and the total number of citations of all its papers. We do not show the scatter plot, only its median, separately for each arXiv category. In both cases the total number of citations grows roughly linearly with the number of authors.* **Right**: *As in the middle panel, but showing on the vertical axis the number of individual citations, which roughly does not depend on the number of authors.*

received by an author grows roughly linearly with the mean number of co-authors. The right panel of Fig. 3 shows that, instead, the total number of individual citations is scale-invariant, namely roughy independent of the number of co-authors. This means that individual citations do not reward nor penalise big collaborations, while citations reward big collaborations.

3.3 PaperRank of authors: results

We now apply the improved metrics described in the previous section to the InSpire database.

The left column of Table 7 lists the authors with the highest number of individual citations. Two factors differentiate citations from individual citations. First, dividing by the number of references (first factor in the denominator in Eq. (12)) counter-acts the inflation in the total number of citations. This factors mildly penalises authors working in sectors (such as hep-ph) where papers have a larger average number of references. Second, dividing by the number of authors (second factor at the denominator in Eq. (12)) has a large impact: members of huge collaborations no longer make the top positions of the list, which becomes dominated by theorists. As discussed in Section 3.2, this happens because working in large collaborations does not allow to recognise individual merit — not because working in large collaborations decreases the average merit. Lists of bottom authors would similarly be dominated by theorists.

The left column of Table 8 shows the top-ranked authors in the InSpire database. The PaperRank of authors identifies some older notable authors who received the prizes plotted in

	InSpire name	All InSpire	InSpire name	After 2000	InSpire name	After 2010
1	Steven.Weinberg.1	235825	E.Witten.1	3670	N.Kidonakis.1	822.3
2	J.S.Schwinger.1	231194	J.M.Maldacena.1	3551	E.Witten.1	794.2
3	R.P.Feynman.1	162394	S.D.Odintsov.1	3027	J.M.Maldacena.1	786.5
4	M.Gell.Mann.1	156387	T.Sjostrand.1	3016	U.G.Meissner.1	785.3
5	C.N.Yang.1	114656	D.T.Son.1	2958	S.Sachdev.1	776.9
6	P.A.M.Dirac.1	111083	T.Padmanabhan.1	2824	S.Hod.1	752.2
7	Abdus.Salam.1	104766	P.Z.Skands.1	2804	D.Stanford.1	715.1
8	E.Witten.1	102753	S.Nojiri.1	2783	X.L.Qi.1	713.0
9	H.A.Bethe.1	102195	V.Springel.1	2701	S.D.Odintsov.1	704.1
10	G.tHooft.1	99645	A.Loeb.1	2594	Muhammad.Sharif.1	701.7
11	E.P.Wigner.1	90966	C.Vafa.1	2497	A.Loeb.1	700.3
12	T.D.Lee.1	86699	Ashoke.Sen.1	2468	S.Capozziello.1	674.2
13	W.Heisenberg.1	81238	G.P.Salam.1	2453	C.de.Rham.1	657.4
14	Stephen.Louis.Adler.1	76687	Martin.Bojowald.1	2373	L.Susskind.1	650.0
15	Yoichiro.Nambu.1	74084	S.Mrenna.1	2352	S.Tsujikawa.1	588.1
16	K.G.Wilson.1	69320	Ernest.Ma.1	2297	Bing.Zhang.1	577.6
17	S.L.Glashow.1	69308	A.C.Fabian.1	2285	D.W.Hooper.1	573.5
18	F.J.Dyson.1	68021	F.Aharonian.1	2189	C.Bambi.1	564.9
19	A.M.Polyakov.1	67138	Wayne.Hu.1	2162	R.B.Mann.1	564.0
20	J.D.Bjorken.1	63817	P.Nason.1	2132	A.Strominger.1	548.2
21	S.W.Hawking.1	62039	E.V.Shuryak.1	2121	E.Oset.1	534.9
22	B.Zumino.1	58680	G.R.Dvali.1	2100	P.Z.Skands.1	526.5
23	S.Mandelstam.1	57268	U.G.Meissner.1	2081	Xiao.Gang.Wen.1	520.4
24	G.Breit.2	52537	S.S.Gubser.1	2003	S.Nojiri.1	514.0
25	V.F.Weisskopf.1	52355	D.E.Kharzeev.1	1982	H.T.Fortune.1	512.1
26	Enrico.Fermi.1	52039	A.A.Tseytlin.1	1946	S.F.King.1	503.2
27	J.R.Oppenheimer.1	51472	M.Cacciari.1	1937	J.R.Ellis.1	500.6
28	David.J.Gross.1	50937	M.Zaldarriaga.1	1911	C.L.Kane.1	500.0
29	Peter.W.Higgs.1	50761	U.W.Heinz.1	1876	R.Myrzakulov.1	497.8
30	J.A.Wheeler.1	50238	Joseph.I.Silk.1	1873	A.Vishwanath.1	494.5

Table 8: *Authors sorted according to the PaperRank of authors,* $R_A = \sum_{\text{papers}} R_p/N_{\text{aut}}$ *for the whole* InSpire *database (left), from year 2000 (middle), and from year 2010 (right).*

front of their name, despite having less citations than modern authors, given the increase in the rate of papers and of citations.

Anyhow, the main interest of our study is not re-discovering Feynman. We want to see if our metrics do a better job than just citation counts in identifying modern authors with a high impact. To achieve this, we set a lower cut-off on the publication year. We restrict the list to papers published 'From 2000' (middle column of Table 8) and 'From 2010' (right column of Table 8).

While switching from citations to individual citations is an obvious improvement, we find that the rank does not improve over individual citations (they are strongly correlated) when restricting to recent papers. As already discussed for papers, about 10-20 years are needed before that the PaperRank becomes a better metrics. On shorter time-scales, the rank and the number of individual citations are strongly correlated, and no significant differences arise; a few authors have a rank significantly higher than their number of individual citations often because they happen to be cited by reviews which fast received a large number of citations.

	InSpire name	All	InSpire name	After 2000	InSpire name	After 2010
1	P.A.M.Dirac.1	292479	E.Witten.1	18748	N.Kidonakis.1	7248
2	Albert.Einstein.1	229323	T.Sjostrand.1	16669	E.Witten.1	5178
3	E.Witten.1	166374	J.M.Maldacena.1	16406	J.M.Maldacena.1	4799
4	Steven.Weinberg.1	147661	V.Springel.1	15267	P.Z.Skands.1	4335
5	G.tHooft.1	109341	P.Z.Skands.1	13705	A.A.Abdo.1	4294
6	J.S.Schwinger.1	98088	S.Mrenna.1	11931	M.Czakon.1	4160
7	Max.Born.1	81451	C.Vafa.1	11840	D.Stanford.1	4160
8	S.W.Hawking.1	79424	G.P.Salam.1	11091	X.L.Qi.1	3970
9	A.M.Polyakov.1	69293	D.T.Son.1	10511	M.Luscher.1	3929
10	R.P.Feynman.1	68094	P.Nason.1	10405	G.Harry.1	3666
11	M.Gell.Mann.1	66393	J.A.M.Vermaseren.1	9195	L.Susskind.1	3322
12	C.N.Yang.1	65056	Alan.D.Martin.1	8994	Alexander.Romanenko.1	3318
13	Enrico.Fermi.1	64907	S.D.M.White.1	8761	Xiao.Gang.Wen.1	3249
14	K.G.Wilson.1	54851	B.R.Webber.1	8673	S.Sachdev.1	3224
15	H.A.Bethe.1	53697	A.Loeb.1	8519	A.Strominger.1	3202
16	T.D.Lee.1	50413	Ashoke.Sen.1	8427	H.T.Janka.1	3108
17	L.Susskind.1	46873	L.E.Hernquist.1	8331	P.Nason.1	3089
18	Abdus.Salam.1	46278	J.Polchinski.1	8193	J.Rojo.1	3083
19	S.L.Glashow.1	45184	M.Cacciari.1	8146	A.Mitov.1	3077
20	R.W.Jackiw.1	44070	S.Frixione.1	8091	G.P.Salam.1	2934
21	F.A.Wilczek.1	43153	L.Susskind.1	8030	C.L.Kane.1	2831
22	H.M.Georgi.1	40178	A.C.Fabian.1	7809	F.Maltoni.1	2775
23	Stephen.Louis.Adler.1	39928	Wayne.Hu.1	7797	A.Vishwanath.1	2739
24	E.P.Wigner.1	39752	A.Strominger.1	7750	U.G.Meissner.1	2702
25	Sidney.R.Coleman.1	39247	M.Luscher.1	7710	J.Polchinski.1	2649
26	David.J.Gross.1	38777	Nathan.J.Berkovits.1	7679	Patrick.Huber.1	2606
27	A.D.Linde.1	38338	Xiao.Gang.Wen.1	7623	C.de.Rham.1	2577
28	J.D.Bjorken.1	38272	U.G.Meissner.1	7613	A.Loeb.1	2521
29	W.Heisenberg.1	36412	A.Y.Kitaev.1	7472	E.Berger.1	2492
30	S.Mandelstam.1	35261	F.Aharonian.1	7438	S.Forte.2	2463

Table 9: *Authors sorted according to their Author Rank $\mathscr{R}_A$ for the whole* InSpire *database (left), from year 2000 (middle), and from year 2010 (right).*

Restricting the sums to the N-th best papers of each author has little effect.

3.4 Author Rank

As outlined in the introduction, the citation matrix between authors (properly normalized) $C_{A'\to A}$ defined in Eq. (6) allows to define an AuthorRank as

$$\mathscr{R}_A = \wp \sum_{A'} \mathscr{R}_{A'} C_{A'\to A} + \alpha(1-\wp)\,, \tag{14}$$

where the second term gives a constant weight to each author, independently from the number and quality of its papers. The network of citations among authors avoids time-directness, up to time-scales comparable to the scientific ages of authors (which is enough for some goals, such as studying how senior authors evaluate the work of younger authors). While formally analogous to the ranking of papers, this ranking of authors is not a model of a physical process, because one reads papers, not authors. The graph corresponding to the matrix $C_{A'\to A}$ contains cycles

	InSpire name	All	InSpire name	After 2000	InSpire name	After 2010
1	E.Witten.1	3428	J.M.Maldacena.1	253.5	J.M.Maldacena.1	68.0
2	Steven.Weinberg.1	2261	T.Sjostrand.1	228.8	D.Stanford.1	50.2
3	G.tHooft.1	1366	P.Z.Skands.1	223.0	S.D.Odintsov.1	49.0
4	S.W.Hawking.1	1217	S.Mrenna.1	207.7	C.de.Rham.1	48.7
5	A.M.Polyakov.1	923.5	E.Witten.1	197.0	E.Witten.1	48.2
6	J.M.Maldacena.1	824.3	S.D.Odintsov.1	192.1	A.Strominger.1	47.9
7	J.S.Schwinger.1	739.7	G.P.Salam.1	181.2	N.Kidonakis.1	47.2
8	T.Sjostrand.1	736.1	S.Nojiri.1	173.1	P.Z.Skands.1	44.7
9	F.A.Wilczek.1	673.7	D.T.Son.1	162.8	M.Luscher.1	39.6
10	L.Susskind.1	658.6	V.Springel.1	157.4	G.P.Salam.1	38.5
11	R.W.Jackiw.1	655.8	T.Padmanabhan.1	156.6	M.Czakon.1	38.2
12	A.D.Linde.1	652.6	M.Cacciari.1	151.7	L.Susskind.1	36.9
13	S.L.Glashow.1	636.8	P.Nason.1	148.5	F.Maltoni.1	34.6
14	N.Seiberg.1	627.8	C.Vafa.1	134.7	S.Nojiri.1	34.0
15	P.A.M.Dirac.1	624.4	Ashoke.Sen.1	125.7	E.P.Verlinde.1	33.5
16	Sidney.R.Coleman.1	622.9	N.Arkani.Hamed.1	119.6	R.C.Myers.1	33.1
17	H.M.Georgi.1	614.4	A.Strominger.1	119.5	S.Tsujikawa.1	32.2
18	K.G.Wilson.1	552.5	V.A.Kostelecky.1	113.1	A.Mitov.1	32.0
19	David.J.Gross.1	543.2	J.Polchinski.1	108.0	G.Soyez.1	30.8
20	C.Vafa.1	521.4	G.Soyez.1	99.7	P.Nason.1	30.6
21	M.Luscher.1	508.0	D.E.Kharzeev.1	97.6	M.Cacciari.1	29.7
22	R.P.Feynman.1	459.9	P.Horava.1	97.3	A.D.Linde.1	28.4
23	A.Strominger.1	450.9	S.S.Gubser.1	96.5	A.A.Abdo.1	28.2
24	R.L.Jaffe.1	441.2	M.Luscher.1	94.6	Olivier.Mattelaer.1	28.0
25	J.Polchinski.1	439.7	S.Tsujikawa.1	93.8	A.De.Felice.1	27.9
26	J.D.Bjorken.1	422.5	L.Susskind.1	93.7	X.L.Qi.1	26.8
27	M.Gell.Mann.1	420.0	S.Frixione.1	92.7	D.Simmons.Duffin.1	26.7
28	C.N.Yang.1	415.9	N.Seiberg.1	90.3	K.Hinterbichler.1	26.4
29	B.Zumino.1	412.8	A.Ashtekar.1	89.4	R.Venugopalan.1	26.4
30	Abdus.Salam.1	410.0	M.A.Stephanov.1	87.5	N.Seiberg.1	26.2

Table 10: *Authors sorted according to their Citation-coin $\not{C}_A$ of Eq. (15) for the whole* InSpire *database (left), from year 2000 (middle), and from year 2010 (right)..*

and also loops on the same node (self-citations): here we use $\wp = 0.9$, such that self-citations cannot boost the AuthorRank by more than one order of magnitude.

The left column of Table 9 shows the all-time AuthorRank. We see the emergence of old authors such as Einstein, which were absent from previous top-rankings because poorly covered and cited in the too recent InSpire database. Of course, the incompleteness of InSpire before $\sim$ 1960 makes results about older authors semi-quantitative.

This issue is avoided in the other columns of Table 9, where we show the AuthorRank recomputed by restricting to recent papers only.

3.5 Removing self-citations and citation 'cartels': the Citation-coin

One of the unsatisfactory aspects of previous metrics is the effect of self-citations.

On short time-scales the PaperRank R and the number of (individual) citations are strongly correlated, so they can be similarly inflated trough self-citations. Only on longer time-scales R becomes a more robust measure: citations from paper p' are weighted by its rank $R_{p'}$, giving

relatively less weight to 'below average' papers that sometimes contain many self-citations. Still, many below average papers can sum up to a significant total rank.

One can optionally count citations from published papers only, ignoring citations from unpublished papers. However this choice discards information on good unpublished papers (some well respected authors do not publish some of their papers).

Removing all self-citations is, by itself, an arbitrary choice. Furthermore it can be implemented in different ways, for example removing citations from co-authors. Removing only citations of an author to itself amounts to set to zero the diagonal elements of the citation matrix $N^{\rm icit}_{A'\to A}$, reducing its N^2 entries to $N(N-1)$.

This does not protect from 'citation cartels'. A second step in this direction consists in removing citations exchanges $A \to A'$ and $A' \to A$ between all pairs of authors A, A'. This amounts to subtract the symmetric part of the $N^{\rm icit}_{A\to A'}$ matrix, reducing its entries to $N(N-1)/2$.

A third step is removing citations exchanges $A \to A'$, $A' \to A''$ and $A'' \to A$ among triplets of authors A, A', A''. A fourth step is removing all quadruplets etc.

A combinatorial computation shows that, after removing all possible 'cartels', only $N-1$ entries remain. They can be described by N numbers ₡_A that sum up to 0. The meaning of ₡_A can be intuitively understood by viewing $N^{\rm icit}_{A'\to A}$ as the total amount 'paid' by A' to A, and $N^{\rm icit}$ as a matrix of transactions. Then the physical quantity unaffected by cartels is the net amount 'owned' by each author A: $\text{₡}_A = \sum_{A'}(N^{\rm icit}_{A'\to A} - N^{\rm icit}_{A\to A'})$. Citations are treated like money: subtracting all possible citation 'cartels' is equivalent to count citations received as positive, and citations given as negative. In doing this we proceed as described above: we actually count individual citations ('icit'): shared between co-authors, and divided by the number of references of each paper. Then the price paid is the total number of papers written, independently of their number of references. The Citation-coin of authors, ₡_A, can be written in terms of the Citation-coin of their papers, ₡_p:

$$\text{₡}_A = \sum_{p\in A} \frac{\text{₡}_p}{N^{\rm aut}_p}\,, \qquad \text{₡}_p = N^{\rm icit}_p - f(t_{\rm end} - t_p) \tag{15}$$

with $f(t_{\rm end} - t_p) = 1$. A paper has $\text{₡}_p < 0$ when it receives a below-average number of individual citations. This is the case for all recent papers, suggesting the introduction of a factor $f(t_{\rm end} - t_p) \leq 1$. All metrics penalise recent papers and authors (namely those active a few years before $t_{\rm end}$, the end-date of the database), because citations accumulate over time. This penalisation is stronger in the Citation-coin than in citation counting, if a paper p is paid at the moment of its appearance t_p. This boundary effect can be compensated by redefining the Citation-coin in such a way that papers are 'paid in instalments', with the same time-scale over which citations accumulate. One can conveniently choose $f(t) \approx 1 - e^{-t/\tau}$, where $\tau \approx 7\,$yr is presently the time-scale over which citations are received by an average paper in the InSpire database.[17] In this way $f(t) \simeq 1$ for old papers, $f(t) < 1$ for young papers, $f(t) \simeq 0$ for very recent papers. As a consequence recent authors do not necessarily have a negative Citation-

[17]This number is obtained considering the whole InSpire and it varies by a factor of around two across the main arXiv categories. As for fractional counting, in the case one wants to fine-tune comparisons within a given sub-field, this number can be tuned to the behavior of that sub-field.

coin, and the lack of knowledge of future citations does not penalise their $\not\!\!C$ more than their number of citations.[18]

Table 10 lists authors according to their Citation-coin $\not\!\!C_A$. Authors that scored highly in previous ranks by writing many papers with low impact have now disappeared, and some of them actually got a negative $\not\!\!C_A$.

It is interesting to compare the Citation-coin to traditional metrics. Unlike the number of (individual) citations, and unlike the number of papers, the $\not\!\!C$ does not reward authors that publish many low impact papers. Like the h-index, the $\not\!\!C$ rewards both quality and quantity, but in a different way. The h-index puts a threshold on the number of citations: as a result papers below don't contribute, and all papers above count equally. On the other hand, excellent papers contribute significantly to the $\not\!\!C$, while below-average papers contribute negatively. In this respect, the Citation-coin can be considered an improvement of the h-index that is both more indicative and less correlated (see Section 3.6) with citation counting. The $\not\!\!C$ also differs from $\langle N_{\rm cit}\rangle$, the number of (possibly individual) citations averaged over all papers of the author under consideration. Unlike the $\not\!\!C$, $\langle N_{\rm cit}\rangle$ only rewards quality: an author can maximise it by writing only very few excellent papers. At the practical level, the $\not\!\!C$ identifies physicists considered most esteemed (see Table 10), while traditional metrics fail (see Table 6).

Finally, from the database we can extract detailed reports about the metrics of each author: $N_{\rm pap}$, $N_{\rm cit}$, h, $N_{\rm ipap}$, $N_{\rm icit}$, R, $\mathscr{R}$, $\not\!\!C$, their time evolution, the scientific age, the percentage of given and received self-citations, the topics studied, the main collaborators, who the author cites most, who cites the author most, etc. Similarly, these informations can be extracted and compared for a group of authors. These detailed reports would be a good target for a future improvement of the InSpire author profiles.

3.6 Author metrics: correlations

The right panel of Fig. 2 shows the correlations among the metrics for authors within the whole InSpire database. The metrics are: number of papers $N_A^{\rm pap}$, number of citations $N_A^{\rm cit}$, h-index squared h_A^2, number of individual papers $N_A^{\rm ipap}$, number of individual citations $N_A^{\rm icit}$, PaperRank of authors R_A, AuthorRank of authors $\mathscr{R}_A$, and Citation-coin $\not\!\!C_A$. We consider the square of the h index since this is known to be almost fully correlated (0.99) with the number of citations (Hirsch, 2005). From the table we see that our metrics for authors differ strongly from the traditional ones, and also mildly differ between them. The main difference arises because of the difference between experimentalists and theorists, so that the metrics become more correlated if restricted within each group. However, the combined effect of dividing by the number of references of the citing paper and the number of co-authors of the cited one makes our proposed bibliometric indicators fairly uncorrelated with the existing ones.

[18] Furthermore, one can define a $\not\!\!C^+$ metric which discards 'negative' papers with $\not\!\!C_p < 0$ and sums the contributions of 'positive' papers.

	Institution	All	Institution	After 2010	hep-ex	hep-ph	hep-th	gr-qc	astro-ph	hep-lat	nucl-th	nucl-ex
1	CERN	25065	CERN	2817	413	652	184	10	49	66	8	84
2	Princeton U.	10774	Fermilab	1346	227	289	3	3	111	23	2	10
3	Harvard U.	10702	Brookhaven	791	77	202	7	0	15	103	54	73
4	SLAC	9913	Perimeter Inst.	751	0	81	434	133	35	2	0	0
5	Princeton, Inst	9699	DESY	705	147	274	70	1	38	10	1	9
6	Fermilab	9536	Princeton, Inst	686	0	41	522	6	90	0	0	0
7	Caltech	8470	Dubna, JINR	639	124	137	55	9	12	6	34	53
8	MIT, LNS	7658	KEK, Tsukuba	629	99	70	37	6	24	45	5	5
9	Brookhaven	7149	SLAC	603	71	177	47	2	72	0	1	5
10	LBL, Berkeley	6353	Caltech	584	40	53	159	93	166	1	2	4
11	Cambridge U.	5586	Beijing, Inst.	581	204	125	18	5	69	4	9	4
12	SUNY, Stony Bro	5371	Princeton U.	554	39	37	232	24	132	0	0	5
13	UC, Berkeley	4838	Munich, Max Pla	525	56	191	123	2	53	12	0	6
14	DESY	4580	LBL, Berkeley	519	57	84	27	1	80	4	28	37
15	Moscow, ITEP	4495	Valencia U., IF	492	52	304	7	14	36	7	17	1
16	Dubna, JINR	4431	Imperial Coll.,	488	80	37	179	25	42	1	1	6
17	Washington U.,	4089	Cambridge U., D	460	0	29	262	79	63	16	3	0
18	Munich, Max Pla	4024	Tokyo U., IPMU	416	27	116	165	15	67	2	1	0
19	Maryland U.	3801	Heidelberg, Max	400	33	221	1	0	68	0	3	9
20	Cornell U., LNS	3786	Stanford U., Ph	383	14	20	215	16	55	0	0	4
21	Los Alamos	3573	INFN, Rome	365	84	68	12	32	61	10	5	17
22	Oxford U.	3570	APC, Paris	359	22	9	57	72	170	0	0	0
23	Chicago U., EFI	3569	UC, Berkeley	358	15	79	70	7	103	4	3	5
24	Columbia U.	3560	Maryland U.	352	24	87	35	49	74	14	6	6
25	Imperial Coll.,	3414	Heidelberg U.	352	63	85	44	4	37	10	18	23
26	Pennsylvania U.	3304	Potsdam, Max Pl	348	0	3	166	132	30	0	1	0
27	UCLA	3200	MIT	348	65	35	37	18	35	4	26	24
28	KEK, Tsukuba	3164	Munich, Tech. U	344	24	196	9	0	23	4	26	12
29	Wisconsin U., M	3101	Washington U.,	338	21	57	48	6	22	67	41	12
30	Stanford U., Ph	3080	Harvard U., Phy	334	15	48	189	2	16	0	0	0
31	Saclay	3054	Jefferson Lab	333	10	119	0	0	0	48	22	41
32	Texas U.	2941	Wisconsin U., M	332	61	102	23	0	96	0	3	6
33	UC, Santa Barba	2920	Zurich, ETH	332	45	66	79	6	31	13	0	5
34	Santa Barbara,	2906	Zurich U.	324	40	190	2	12	56	0	0	4
35	Yale U.	2887	Frascati	324	81	46	13	3	3	6	0	22
36	Heidelberg U.	2847	Oxford U.	315	75	35	14	12	108	0	3	2
37	Harvard-Smithso	2633	Durham U., IPPP	314	2	267	14	0	18	0	0	0
38	Lebedev Inst.	2609	NIKHEF, Amsterd	312	61	114	18	21	23	0	0	18
39	Illinois U., Ur	2586	SUNY, Stony Bro	310	35	56	66	0	29	7	57	23
40	Argonne	2559	UC, Santa Barba	306	23	17	180	21	28	6	0	3
41	Kyoto U.	2445	Moscow, ITEP	305	55	72	66	2	6	13	7	33
42	Rutgers U., Pis	2441	Frankfurt U.	304	4	85	7	20	17	49	91	12
43	Tokyo U.	2430	Kyoto U., Yukaw	302	0	36	132	32	44	16	25	0
44	Durham U.	2397	UCLA	301	19	33	105	2	77	1	3	15
45	St. Petersburg,	2385	INFN, Pisa	296	63	45	25	16	44	15	13	3
46	Cambridge U., D	2343	Tokyo U.	294	19	78	43	9	25	23	22	15
47	Moscow, INR	2234	Ohio State U.	286	28	61	24	0	40	12	89	10
48	ICTP, Trieste	2183	Columbia U.	285	19	37	48	5	91	31	10	9
49	Bohr Inst.	2163	McGill U.	284	14	87	80	8	46	1	43	0
50	Minnesota U.	2139	Kyoto U.	282	20	36	77	23	51	12	23	6
51	Utrecht U.	2117	INFN, Turin	281	41	88	27	4	27	6	7	36
52	Frascati	2097	Bonn U.	279	55	93	27	0	16	11	10	8
53	Novosibirsk, IY	2096	Cracow, INP	278	47	134	0	0	6	0	57	12
54	Landau Inst.	2034	Harvard U.	278	5	35	177	6	33	0	0	0
55	Michigan State	2033	Madrid, IFT	274	1	105	96	9	38	8	3	0
56	Ohio State U.	2021	Los Alamos	271	20	78	3	3	26	18	33	22
57	Michigan U.	1963	UC, Irvine	267	38	100	6	0	83	0	0	0
58	Rutherford	1951	Orsay, LPT	265	1	179	30	26	3	13	0	0
59	Weizmann Inst.	1950	Tata Inst.	264	22	36	66	40	10	20	20	6
60	Rome U.	1897	Weizmann Inst.	258	16	55	80	3	29	1	1	6
61	MIT	1880	Southampton U.	253	1	134	48	26	13	25	0	0
62	Munich, Tech. U	1865	Stanford U., IT	253	1	45	180	1	14	0	0	0
63	Serpukhov, IHEP	1855	Pennsylvania U.	252	22	5	111	3	64	0	0	2
64	Penn State U.	1813	Bohr Inst.	252	18	52	103	6	33	14	3	12
65	Orsay, LPT	1804	Aachen, Tech. H	247	66	73	1	1	49	0	0	8
66	Ecole Normale S	1791	Yale U.	246	29	16	32	1	39	5	26	25
67	Hamburg U.	1777	Moscow, INR	241	39	43	23	4	47	0	3	18
68	Syracuse U.	1756	Darmstadt, GSI	240	7	36	1	0	13	5	42	26
69	Indiana U.	1755	Saclay, SPhT	237	0	102	75	3	19	0	27	2
70	UC, San Diego	1717	INFN, Padua	237	49	58	24	6	37	0	1	20
71	Rochester U.	1714	Argonne	236	33	83	2	0	12	5	15	11
72	Carnegie Mellon	1687	Manchester U.	235	65	73	12	0	50	0	6	2
73	Beijing, Inst.	1683	Johns Hopkins U	233	16	64	19	3	122	0	0	2
74	Garching, Max P	1665	INFN, Trieste	230	21	45	45	23	43	0	0	7
75	Chicago U.	1656	British Columbi	229	18	15	79	17	68	0	1	2
76	Frankfurt U.	1644	Michigan State	226	27	106	1	0	17	6	24	12
77	Tel Aviv U.	1604	Texas A-M	225	31	43	50	2	25	0	29	18
78	Texas A-M	1596	Humboldt U., Be	225	15	67	67	2	13	46	0	0
79	Zurich, ETH	1587	Rome U.	224	22	47	13	37	61	6	4	4
80	Bern U.	1584	MIT, Cambridge,	218	0	93	85	0	14	16	4	0

Table 11: *Institutions listed according to their contribution to fundamental physics (as defined by* INSPIRE*) quantified as the number of individual citations received by their affiliates as defined in Eq. (16). Left: all time. Right: from 2010, and split within arXiv categories. The top ten in each category are highlighted.*

4 Rankings groups

Our metrics respect sum rules and thereby allow to define metrics for groups by simply summing over their members. Furthermore, the main property of the Citation-coin holds not only for authors of a set of papers, but for any group: it cannot be increased trough internal citations. In order to show illustrative results we mostly use the number of individual citations as the metric to rank groups, since individual citations have the benefit of being very fast to compute. Of course, one could equally apply any other metric discussed before.

4.1 Ranking institutions

We share the number of citations received by each paper between the institutions I of its authors with weights p_I that sum up to 1. The weights are computed by first sharing equally each paper between its $N_{\rm aut}$ authors, and next between the affiliations of each author. Thereby each author A contributes to institute I as $1/(N_A^{\rm aff} N_p^{\rm aut})$, where $N_A^{\rm aff}$ is the number of affiliations of author A, that include institute I. When some affiliation is missed by InSpire, we renormalize the p_I such that they sum up to 1. Next, we sum over all papers p (optionally restricting to recent papers, if one wants to evaluate the present situation, rather than the all-time record). In formulæ, the number of individual citations received by institute I is:

$$N_I^{\rm icit} = \sum_p p_{pI} N_p^{\rm icit} \,, \qquad p_{pI} = \frac{1}{N_p^{\rm aut}} \sum_{A \in I,p} \frac{1}{N_A^{\rm aff}} \,. \tag{16}$$

In the left column of Table 11 we list the institutes that most contributed to fundamental physics, according to the whole InSpire database. In order to focus on the present situation, in the second column we restrict to recent papers, written from 2010. The top positions are occupied by research institutions rather than by teaching institutions. In the right columns of Table 11 we show the contributions within each main arXiv category: the best institutions strongly depend on the sub-field of interest. This means that generic rankings (e.g. at the faculty level) are not much useful for authors interested in finding the most active institutions within their specific sub-fields.

Concerning the time evolution, we compute the percentage of individual citations received by papers written within any given year by authors affiliated to each institute, and fully fractionalized according to Eq. (16). Fig. 4a shows how the percentage impact of some main institutes evolved with time. As papers published recently will accumulate most of their citations in the future, in Fig. 4b we repeat the analysis in terms of fractionally counted papers. This indicator provides a less significant proxy for scientific merit than citations, but its value is immediately available, and helps to interpret the evolution in citations.

The black curve in Fig. 7 shows the time-dependence of the contribution of the top institution, CERN. It reached a maximum around 1965 (12% of world-wide individual citations to 1965 papers have been given to CERN authors) and declined stabilising to $\approx 2\%$. All main historical institutions show a similar trend, due to the fact that, until 1970, fundamental physics was concentrated in a few institutions, and later became more distributed (especially

	Institution	$N_{\rm iaut}$	$N_{\rm icit}$	PaperRank R	AuthorRank $\mathscr{R}$
1	Princeton, Inst. Adv	32	2.1 %	2.2 %	2.4 %
2	CERN	896	1.7 %	2.0 %	2.7 %
3	Fermilab	435	1.4 %	1.3 %	1.8 %
4	Texas U.	46	1.0 %	3.6 %	1.6 %
5	Caltech	90	0.89 %	3.6 %	1.6 %
6	Brookhaven	218	0.83 %	0.97 %	0.94 %
7	Cambridge U., DAMTP	49	0.78 %	0.96 %	0.79 %
8	SLAC	141	0.73 %	0.83 %	0.90 %
9	Dubna, JINR	312	0.62 %	0.54 %	0.51 %
10	DESY	276	0.59 %	0.46 %	0.57 %
11	Princeton U.	74	0.59 %	0.74 %	0.71 %
12	Harvard U., Phys. De	35	0.58 %	0.89 %	0.74 %
13	Maryland U.	72	0.55 %	0.45 %	0.46 %
14	Imperial Coll., Lond	121	0.54 %	0.56 %	0.55 %
15	Perimeter Inst. Theo	70	0.53 %	0.38 %	0.41 %
16	IPhT, Saclay	40	0.52 %	0.53 %	0.50 %
17	Stanford U., ITP	22	0.51 %	0.69 %	0.62 %
18	LBL, Berkeley	116	0.49 %	0.68 %	0.60 %
19	Munich, Max Planck I	124	0.47 %	0.36 %	0.44 %
20	Chicago U., EFI	41	0.46 %	0.60 %	0.54 %

Table 12: *Institutes listed according to all-time bibliometric ranks of their last-year affiliates. For each institute, we show the number of individual authors active in the last year, and the sum of their biblio-metric indicators (percentage of the world total): number of individual citations, PaperRank of authors, and AuthorRank of authors.*

in theoretical physics). Half of the impact, quantified using individual citations, was produced in the 12 top institutions in 1970, in 22 in 1980, 42 in 1990, 80 in 2000 and 160 now. As a consequence the relative impact of the top institutions declined. Among the new institutions, Perimeter and IPMU reached very high positions.

The list in Table 11 highlights the institutes with the most productive authors in recent times (often young authors). The list in Table 12 highlights institutes with the most productive affiliates, as instead quantified by their all-time biblio-metric rankings. Table 12 is produced as follows. Since no list of present affiliates is available, we use the declared affiliations of authors that wrote at least one paper in the last year, 2020. Authors with $N_{\rm aff}$ affiliations are assigned with fraction $1/N_{\rm aff}$ to each affiliation. When this number differs in recent papers, we average over them respecting sum rules: each authors is affiliated to various institutions with percentages that sum up to one. The average suppresses minor mistakes/missing data in InSpire. For each institute I, we obtain a list of active affiliates with their percentages. Summing over these percentages we determine the number of 'individual authors' $N_{\rm iaut}$ affiliated to each institution, shown in the 3rd column of Table 12. Next, summing over all affiliates using the same weights, we compute the total biblio-metric ranking of all authors in each institute. In column 4 we show the all-time number of individual citations, in column 5 the PaperRank (Section 2.1) and in column the AuthorRank (Section 3.4). The latter 3 columns actually show the world percentage of each metric in the various institutes: about 2% of researchers that most

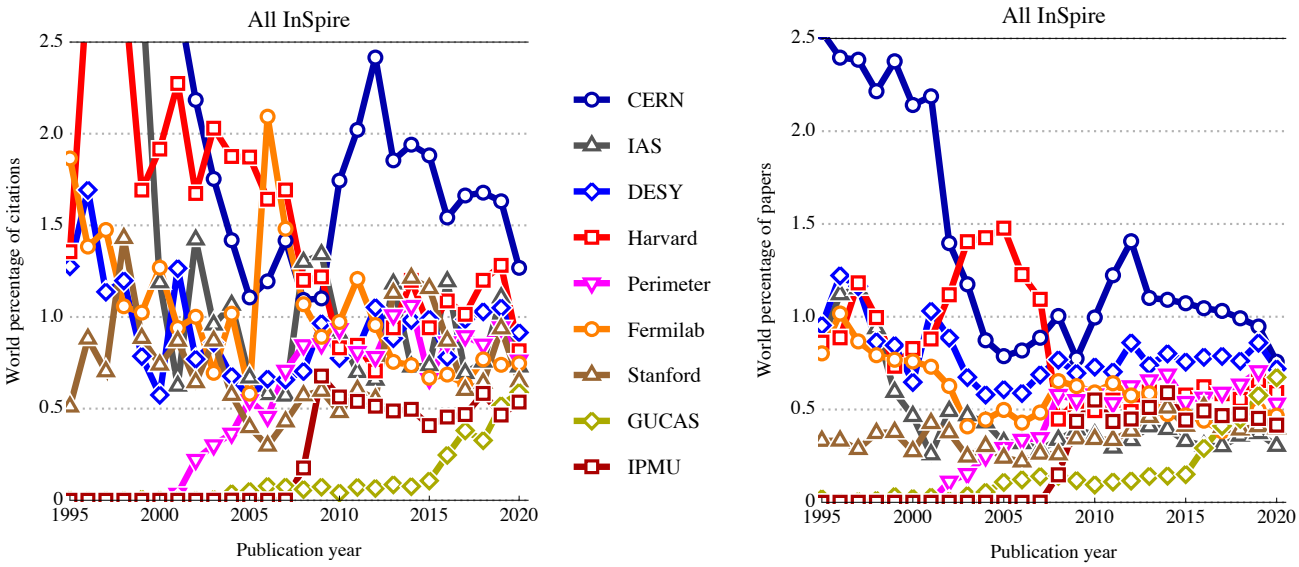

Figure 4: *Percentage impact of some main institutes according to citations (left) and papers (right).*

contributed to fundamental physics can be found at IAS in Princeton, or at CERN.

4.2 Ranking towns

Sometimes multiple institutes are located nearby, and what matters is their total. We group together institutes closer than about 30 km. In Fig. 5 we show a map with the places that mostly contributed to fundamental physics: each contribution is plotted as a circle with area proportional to the number of individual citations received by their papers written from year 2010, and color proportional to the contribution to experiment (green), theory (red), astro-cosmology (blue) respectively. We focus on a relatively recent period, such that the map photographs the present situation.

Similar maps can be computed for any given sub-topic or region. For instance, Fig. 6 shows the same map separated according to papers published within the main arXiv categories, and restricted to Europe.

4.3 Ranking countries and continents

We rank a country or continent C by summing the ranks of all institutes located in C. We apply this to the number of individual citations:

$$N_C^{\text{icit}} = \sum_{I \in C} N_I^{\text{icit}} \,. \tag{17}$$

In the left panel of Fig. 7 we show the time evolution of the impact of papers written in main representative countries within each year. The impact is quantified as the percentage of the

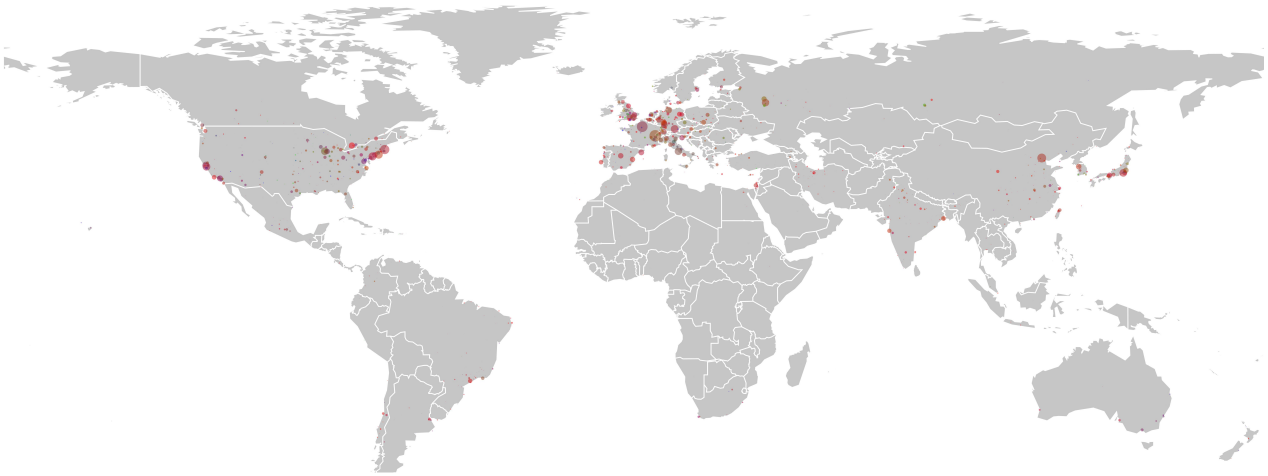

Figure 5: *Institutes that contributed to fundamental physics plotted as dots with area proportional to the number of individual citations received by their papers written from year 2010. We group institutes closer than* 30 km. *The amount of green, red, blue in the color is proportional to the contribution to experiment, theory, astro-cosmology respectively.*

world total, in order to factor out the reduced number of citations of recent papers. USA is the main country, but declining (from 70% around 1950 to 25% now); European countries are now stable or slightly growing; China is growing. In the right panels of Fig. 7 the time evolution of the percentage contribution of each country is shown separately, after the advent of arXiv, within the main fields: experiment, theory, astro-cosmology.

Figure 8 shows the analogous plot for continents. We see that European physics suffered a big decline after WW2, and returned to be the main actor only around 2000. The decline of Asia around 1985 is due to the fall of Soviet Union (Mathematica geographic tools assign all Russia to Asia); the present rise of Asia is mostly due to China.

Figure 9 shows the ratio between the number of individual citations and the population of countries, while Fig. 10 shows the ratio between the number of individual citations and the gross domestic product.

4.4 Ranking journals

When a paper is published on some journal (refereed or not), INSPIRE provides this information. Table 13 lists journals according to the number of individual citations received by all papers they published in fundamental physics, as included in INSPIRE. We separately show these data for all INSPIRE, and restricting to articles published from 2000. Figure 11 shows the time evolution of the percentage number of individual citations received by all papers on selected journals. We see that internet brought a revolution around 2000: the decline of NPB and PLB and the emergence of JHEP and Astrophysics J. as preferred journals.

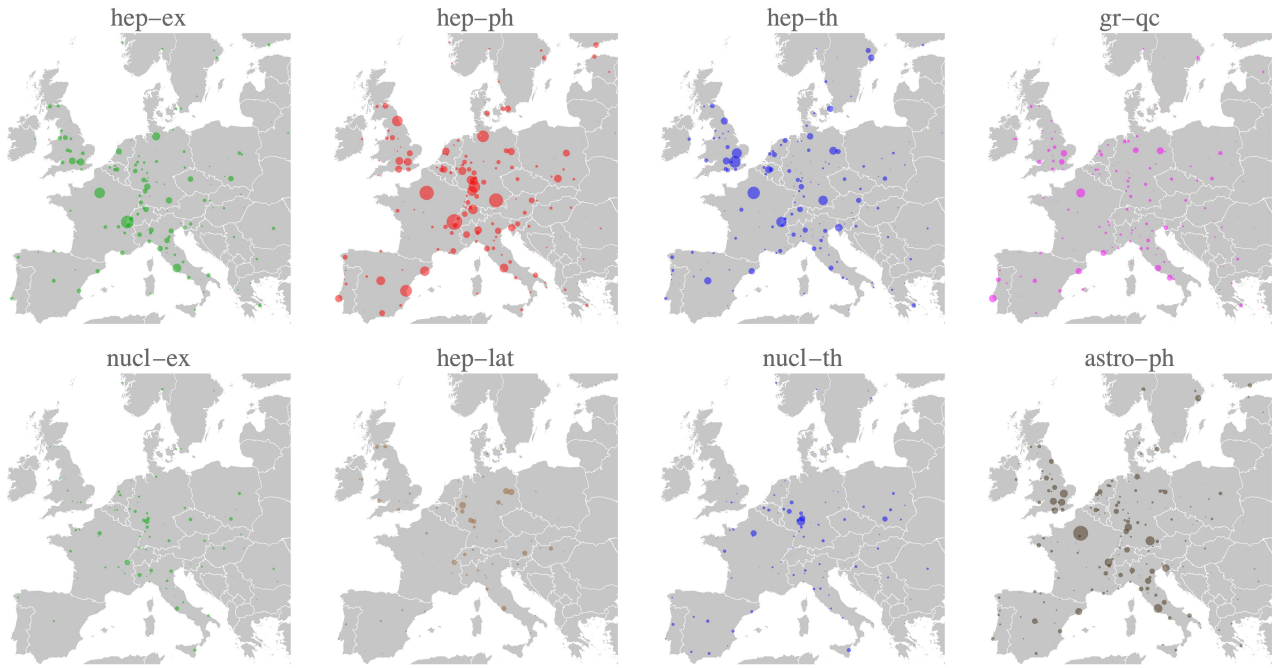

Figure 6: *As in Fig. 5, restricted to Europe, and showing separately the contributions within main arXiv representative categories.*

The 3rd column of Table 13 shows a direct measure of 'quality': the average number of individual citations per paper, which roughly corresponds to what is known as 'impact factor'. According to this measure, the top-journals are those that publish reviews.

The 4th column shows a measure of both 'quality' and 'quantity': the Citation-coin ₡ of the journal (difference between the number of individual citations and the number of published papers). The top journal according to this measure is PRL. Journals that publish reviews score well, but are limited by their restricted scope. The Citation-coin is negative for journals that publish many papers that do not attract much interest, in particular those that publish conference proceedings.

5 Conclusions

We applied improved bibliometric indicators to the InSpire database, which covers fundamental physics mostly after 1970. Figure 13 shows some main trends: growing rate of papers, growing number of authors and of references per paper. Figure 15 shows the health status of the main sub-fields.

The metrics that we explored are:

- The *number of individual citations* N_{icit} — defined as the number of citations divided by the numbers of authors of the cited paper and of references of citing papers — that

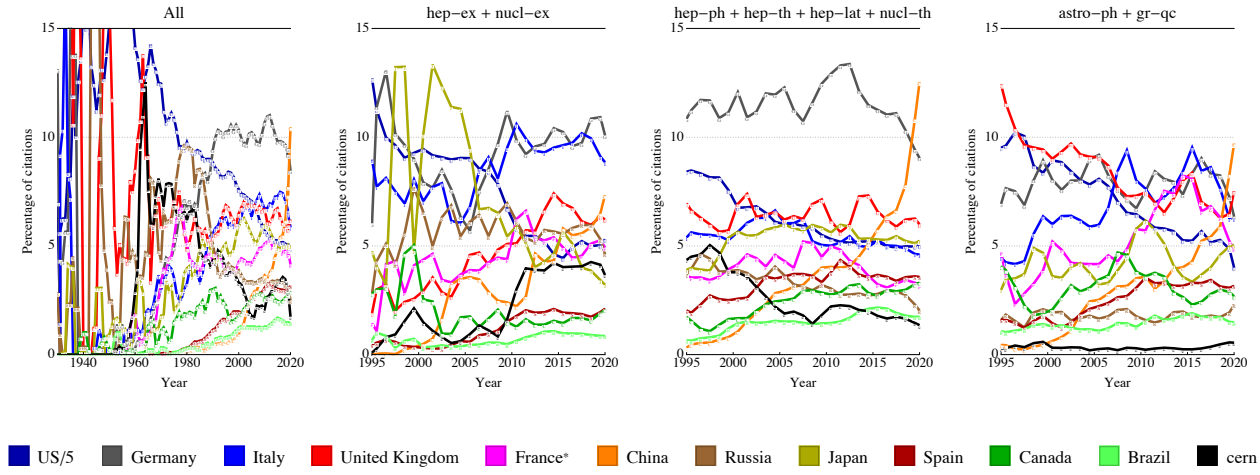

Figure 7: *Percentage impact of representative countries. For ease of visualisation, the USA contribution has been multiplied by 1/5. The ∗ on France means that CERN is plotted separately.*

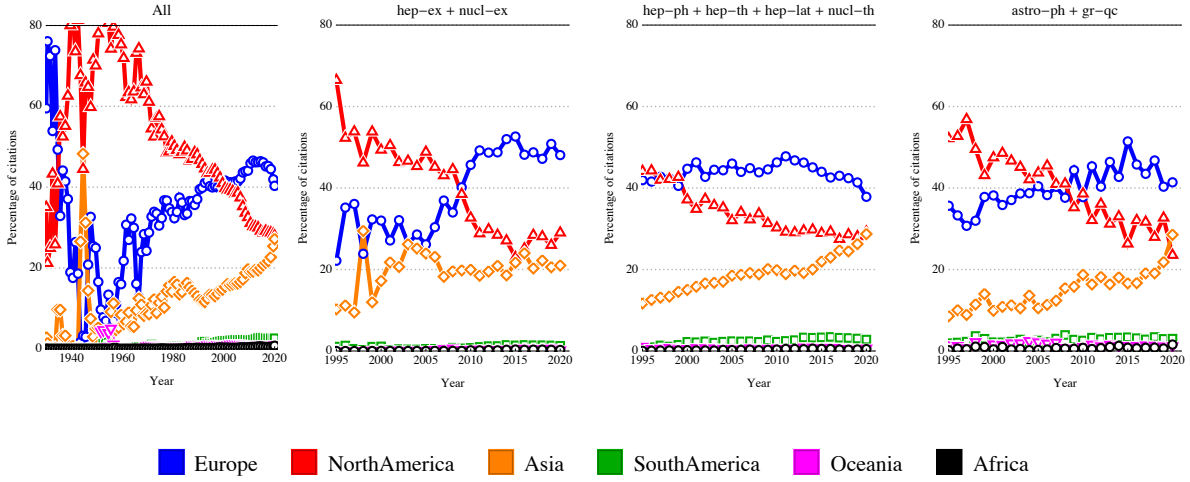

Figure 8: *Percentage impact by continent.*

compensates for the recent inflationary trends towards more authors and references.

- The *PaperRank* R, which applies the PageRank algorithm to the citation network among papers and that approximates a physical observable: how many times a paper is read.
- The *AuthorRank* $\mathscr{R}$ which applies the PageRank algorithm to the individual citation network among authors.
- The *Citation-coin* $\not\mathbb{C}$ equal to the difference between the number of received and given individual citations. Since a naive definition of this indicator, that cannot be increased

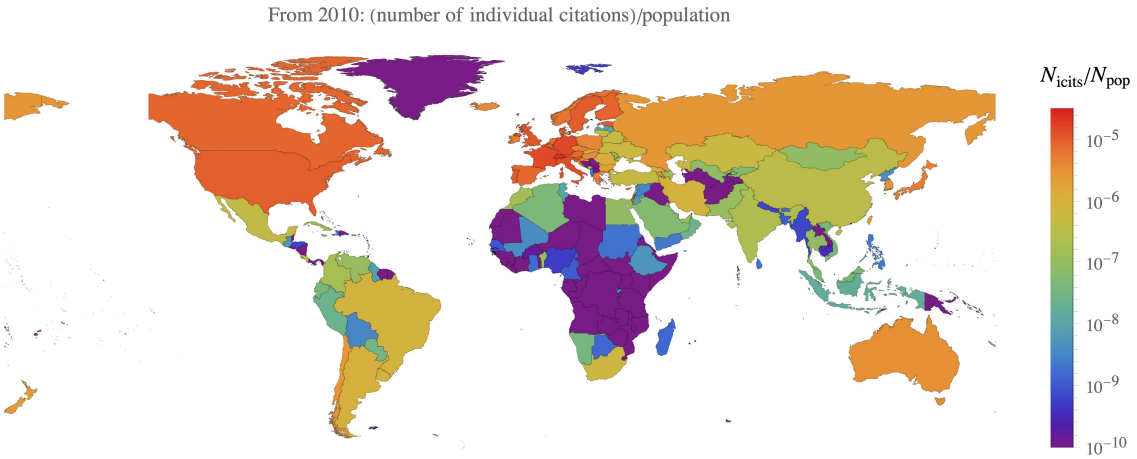

Figure 9: *Individual citations per country from 2010 divided by population. The top countries are Switzerland, Germany, France, Italy, Slovenia.*

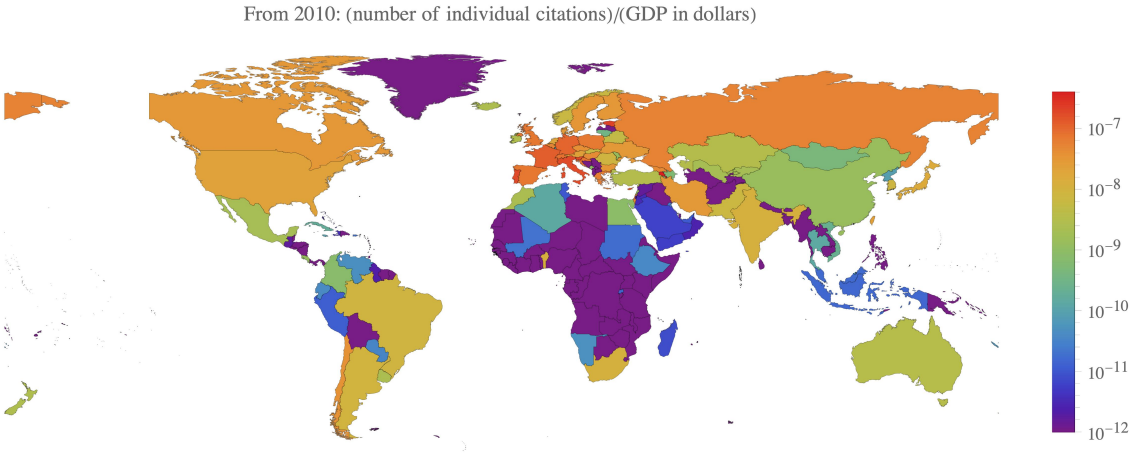

Figure 10: *Individual citations per country from 2010 divided by domestic gross product in dollars. The top countries are Armenia, Estonia, Slovenia, Portugal, Italy.*

by self-citations and circular citations, strongly penalizes young authors, we propose an improvement that takes into account the average time in which citations are received within the community.

An important feature of all these metrics is that they can be computed in practice.

In Section 2 we apply these metrics to papers. Table 1 shows the traditional list of all-time *top-cited* (highest number of citations) papers. It can be compared with Table 2, which shows *top-ranked* papers (papers with highest PaperRank, namely citations weighted proportionally to the rank R of citing papers): the PaperRank retrieves old famous papers with relatively few citations. When applied to the time-ordered citation network, the PageRank reduces to a (weighted) counting of citations-of-citations described in Section 2.3. Thereby, when restricted to recent papers, the PaperRank is dominated by the number of individual citations. Next, Table 3 shows *top-referred* papers, where citations are weighted proportionally to the all-time

	Journal, all INSPIRE	$N_{\rm icit}$	$N_{\rm icit}/N_{\rm pap}$	$\not\Phi$	Journal, after 2010	$N_{\rm icit}$	$N_{\rm icit}/N_{\rm pap}$	$\not\Phi$
1	Phys.Rev.D	113601	1.6	41080	Phys.Rev.D	19288	1.0	656
2	Phys.Lett.B	80214	1.5	25748	JHEP	13942	1.3	3179
3	Phys.Rev.lett.	67864	2.9	44359	Phys.Rev.lett.	9488	2.4	5517
4	Nucl.Phys.B	56150	2.3	31339	Phys.Lett.B	6008	1.3	1261
5	Astrophys.J.	38960	1.3	9920	Astrophys.J.	5929	0.6	−3700
6	Phys.Rev.C	30386	0.9	−4071	Phys.Rev.C	5666	0.9	−371
7	JHEP	30044	1.5	10024	Mon.Not.Roy.Astron.S	5386	0.5	−5091
8	Nucl.Phys.A	21518	0.7	−9310	Eur.Phys.J.C	4376	1.4	1216
9	Mon.Not.Roy.Astron.S	20594	0.9	−2525	JCAP	3174	1.	−105
10	Nucl.Instrum.Meth.A	17592	1.0	380	Astron.Astrophys.	2647	0.5	−3068
11	Phys.Rev.	17351	2.4	10185	Nucl.Instrum.Meth.A	2291	0.7	−898
12	Astron.Astrophys.	14323	0.8	−4762	Class.Quant.Grav.	1896	0.8	−497
13	Astrophys.J.Lett.	10703	1.1	1351	JINST	1890	0.9	−268
14	Eur.Phys.J.C	9964	1.4	2801	Astrophys.J.Lett.	1661	0.7	−683
15	Phys.Rept.	9788	7.4	8462	Pos	1327	0.1	−9215
16	Class.Quant.Grav.	8822	0.9	−731	Nucl.Phys.B	1321	0.8	−327
17	Commun.Math.Phys.	7792	1.9	3658	J.Phys.Conf.Ser.	1183	0.2	−6450
18	Annals Phys.	7765	2.2	4169	J.Phys.G	1046	1.	−44
19	Rev.Mod.Phys.	7235	4.8	5732	Phys.Rev.B	961	0.5	−967
20	Z.Phys.C	6939	1.4	1820	Astrophys.J.Suppl.	903	1.7	386
21	J.Math.Phys.	5717	0.7	−2289	Nucl.Phys.A	847	0.5	−1028
22	Astron.J.	5581	1.1	710	Nature	737	2.9	484
23	JCAP	5331	1.1	509	Eur.Phys.J.A	729	0.7	−252
24	Astrophys.J.Suppl.	4824	2.9	3148	Astropart.Phys.	717	1.4	196
25	Nature	4406	3.0	2945	Chin.Phys.C	706	0.7	−316
26	Prog.Theor.Phys.	4385	0.6	−2697	Phys.Rev.ST Accel.Be	672	0.9	−73
27	Int.J.Mod.Phys.A	4374	0.5	−4459	Comput.Phys.Commun.	630	2.0	314
28	Yad.Fiz.	4072	0.4	−5218	Int.J.Mod.Phys.A	622	0.4	−1152
29	Comput.Phys.Commun.	3864	2.6	2373	Phys.Rept.	584	6.2	490
30	JINST	3816	1.5	1330	Science	548	3.4	387

Table 13: *Number of individual citations ($N_{\rm icit} = \sum N_{\rm cit}/N_{\rm ref}$), average number of individual citations per paper ($N_{\rm icit}/N_{\rm pap}$) and Citation-coin ($\not\Phi = N_{\rm icit} - N_{\rm pap}$) for some top journals. The analysis is restricted to fundamental physics as included in the* INSPIRE *database. Left: all time. Right: only papers published from year 2010.*

AuthorRanks $\mathscr{R}_A$ of citing authors. The AuthorRank identifies the recent papers that most attract the attention of notable older authors. Finally the left panel of Fig. 2 shows the correlations among paper metrics, showing how our proposed metrics are fairly independent from the traditional ones and among each others.

In Section 3 we apply the new metrics to authors. Traditional metrics shown in Table 6 are dominated by experimentalists who write more than 100 papers per year in collaborations with more than 2000 authors. Considering instead the number of individual citations, the list in Table 7 becomes dominated by theorists, especially those very active in relatively recent times. Restricting to recent papers, the list includes some authors from fields that tend produce many publications (tens per author per year). Less surprisingly, the list includes authors who produce useful tools for collider experiments, which are presently very active. Ranking authors trough their PaperRank, the all-time list in Table 8 is dominated by theorists (such as Weinberg,

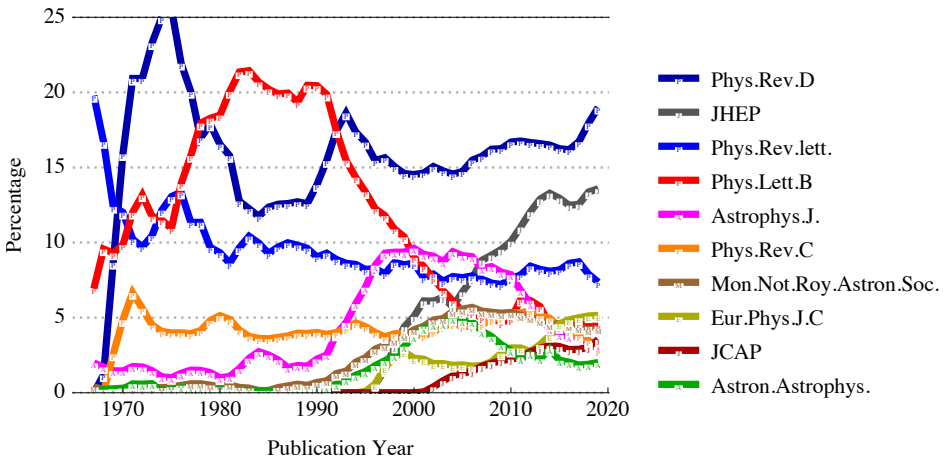

Figure 11: *Time evolution of the fraction of individual citations received by papers published on some notable journals.*

Schwinger, Feynman, Gell Mann) that produced seminal papers after INSPIRE started, despite the overall rate of papers and citations was a factor of few smaller than now (Fig. 14). On the other hand, the recent-time list does not change significantly: the PaperRank is strongly correlated to the number of individual citations, becoming a better metric only on longer time-scales.

Due to this limitation, we developed the AuthorRank. Table 9 shows the result: authors such as Dirac and Einstein now appear in the top of the list, despite having few papers with few citations. The right columns of Table 9 shows recent authors listed weighting citations according to the all-time AuthorRank of citing authors.

Table 10 lists authors according to their Citation-coin $\not{C}$: this metric rewards authors who attract the interest of others by writing above-average papers, and penalising those that write many below-average papers (or many recent papers, as papers need decades to be recognized in terms of number of citations). The right columns of Table 10 again restrict the list to recent times.

All our lists of authors also show the medals and prizes received by the various authors: this shows that the non-traditional metrics agree much better than traditional metrics with the opinions of the various panels (of course an unknown correlation between bibliometrics and prize awards may still be present).

The right panel of Fig. 2 shows the correlations among indicators for authors. The metrics we propose are fairly uncorrelated with traditional metrics and among each other.

Our metrics respect sum rules (their total is not inflated adding more authors or more references) and are intensive: this means that groups can be ranked summing over their members. In Section 4.1 we discussed the institutions that contain the authors that most contributed. In Section 4.2 we grouped nearby institutes, providing maps of towns most active in fundamental physics in different subfields. The same is done in Section 4.3 for countries and continents: in view of the large statistics we also show the time evolution of their percentage impact. Finally,

in Section 4.4 we compute which journals publish the most impactful results in fundamental physics, again showing the time evolution.

The different metrics that we propose give different informations on each author, providing together a more complete view. PAPERSCAPE[19] extracts information from arXiv and provides a very useful visualisation of the citation graph among papers, and of the contribution of some authors (those with unique names). It would be interesting to run the open-source PAPERSCAPE code on the graph of individual citations among papers and authors extracted from the INSPIRE database. Our indices could also be implemented in databases that index citations, such as INSPIRE, in order to offer authors (institutes/journals/group) profiles with a larger variety of information, able to give at one glance a much deeper and wider panoramic of each author (institute/journal/group).

Several of our results with complete tables are available at the PhysRank webpage[20].

Disclaimer Technical details and limitations are described in the appendix. We repeat the main caveats of our analysis: 'fundamental physics' here means 'as included in the INSPIRE database'; we do not correct for mistakes in INSPIRE (anyhow more accurate than commercial databases). Omissions should be addressed to feedback@inspirehep.net or trough the on-line forms on INSPIRE; we will update our results in some future. We just computed and showed results, avoiding comments. We hope that authors of any field, journals boards, members of institutes, towns, countries, and continents will understand that we cannot repeat all caveats in all results.

Note added: 2021 update

All data have been updated to 2021/1/1. This is especially interesting, as many activities switched to on-line mode in view of the covid-19 pandemic started during early 2020. Its effect is in progress and will be better evaluated in the future. So far, Fig. 14 shows no noticeable impact in the number of publications and related bibliometric indicators within the main sub-fields. Similarly, Fig. 15 shows no noticeable impact on the number of authors who left the various sub-fields, nor in the number of new authors. Fig. 7 indicates an accelerated growth of publications from authors with Chinese affiliations relative to others. Fig. 8a shows that Asia overcome NorthAmerica and that the European contribution, while still larger, declined significantly. This is reflected in Fig. 4, that shows the time evolution of the percentage impact of some main institutions.

Acknowledgments

We thank Roberto Franceschini, Christoffer Petersson, and Paolo Rossi for discussions that stimulated this work. We thank Roberto Franceschini for participation in the first stage of this project. We are

[19] http://paperscape.org.
[20] http://rtorre.web.cern.ch/rtorre/PhysRank.

grateful to the InSpire team, and especially to Jacopo Notarstefano, for support and help with the InSpire database.

A The InSpire and arXiv databases

The open-source InSpire bibliographic database[21] covers fundamental physics world-wide. InSpire presently contains about $1.4 \cdot 10^6$ papers, $4 \cdot 10^7$ references, 10^5 authors, 10^4 institutions, and $2 \cdot 10^3$ journals. InSpire started around 1965, but it also contains some notable older papers. InSpire maps papers, authors, collaborations, institutes (affiliations), and journals to record IDs (integer numbers) thereby addressing the problem of name disambiguation (Golosovsky and Solomon, 2017, Martin Montull, 2011).

Starting from 1995, preprints for most of the papers contained in InSpire are available through arXiv.org,[22] which also covers fields beyond fundamental physics, so that not all of the arXiv database is included into the InSpire one. The arXiv also provides a classification in terms of categories, some of which contain sub-classes. The arXiv categories and the number of papers in each of them are shown in the left histogram in Fig. 12.[23] The right histogram shows the fraction of papers in the various categories included in InSpire.

We also often show results for the main arXiv categories inside InSpire, defined as the arXiv categories with more than 10^3 papers, and with a fraction included in InSpire larger than 50%. These are: hep-ex (high-energy experiment), hep-ph (high energy theory/phenomenology), hep-th (high energy theory), astro-ph (astrophysics and cosmology), hep-lat (lattice field theory), nucl-ex (nuclear experiment), nucl-th (nuclear theory), gr-qc (general relativity and quantum cosmology). Details of the dataset we consider and technical issues about the InSpire and arXiv databases are discussed in Appendix A.2.

A.1 Main trends in the fundamental physics literature

The left panel of Fig. 13 shows the time evolution of some main factors: number of papers per year (which increased by 5%/yr); average number of references per paper (increased by 3%/yr); number of citations per paper (roughly constant, taking into account that recent papers, published in the past ≈ 15 years, necessarily received less citations); number of authors per paper (increased from few to tens). We also see that most citations go to published papers. In the right panel of Fig. 13 we see that the average age of references is increasing: before 1980 authors usually mostly cited recent papers. Contemporary papers cite references published Δt earlier with a distribution $\exp(-\Delta t/\tau)$ with $\tau \approx 11\,\mathrm{yr}$, see also Walker, Xie, Yan, and Maslov (2007) and Wang, Song, and Barabási (2013).

Figure 14 shows the same trends within the main arXiv representative categories, showing that papers with an increasingly large number of authors lie in experimental categories (hep-ex,

[21] https://inspirehep.net.

[22] https://arxiv.org.

[23] The full list of sub-classes can be found in the arXiv API reference manual (https://arxiv.org/help/api/user-manual).

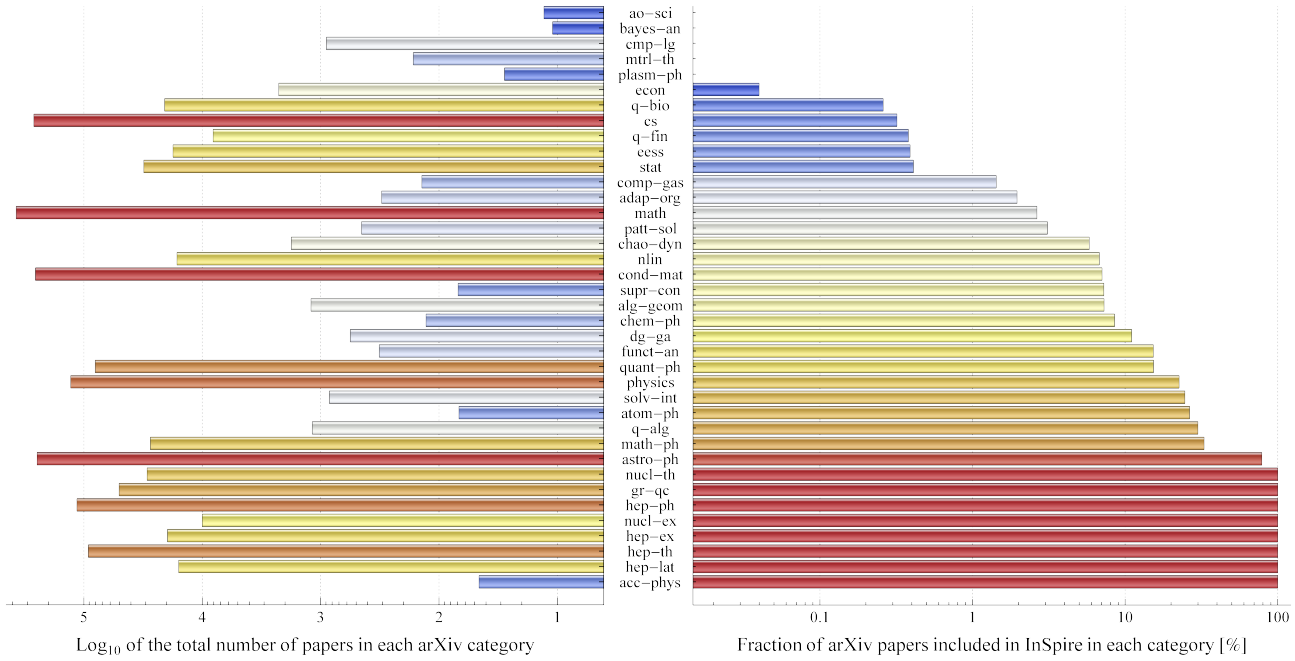

Figure 12: *Total number of papers in arXiv and fraction included in* INSPIRE *by gran categories. The number of papers in each category is computed considering only the main category and ignoring cross-list, so that no paper can belong to more than one gran category.*

nucl-ex, astro-ph), while papers in other fields keep having, on average, $2-3$ authors.

Figure 15 shows the authors' 'birth' and 'death' rates within the main arXiv categories as function of time: in green the percentage of authors who published in year y but not in year $y-1$; in red the percentage of authors who published in year $y-1$ but not in year y. The balance is stable, with a significant growth of hep-ex when the Large Hadron Collider (LHC) started, and of astro-ph until 2010.

Figure 16 shows the distributions of citations and individual citations for the whole INSPIRE database. Citations do not follow a well behaved probability distribution due to the presence, in the whole INSPIRE database, of communities with very different average number of authors. This effect is clearer when going to the individual citations, that apply fractional counting to the citations. Indeed, individual citations are reasonably well described by a log-normal distribution, as already observed by Thelwall and Wilson (2014), with mean and standard deviation both of order one. This shows how individual citations are well described by a multiplicative stochastic process, allowing to combine and compare more heterogeneous subfields.

In order to give an idea of how impact, as defined from citations and individual citations, is distributed within the community, we computed the Gini index of these distributions. The Gini index is used in economy as a measure of inequality of wealth. Just to give an idea of its meaning, typical occidental countries presently have Gini coefficients between 0.2 and 0.4.

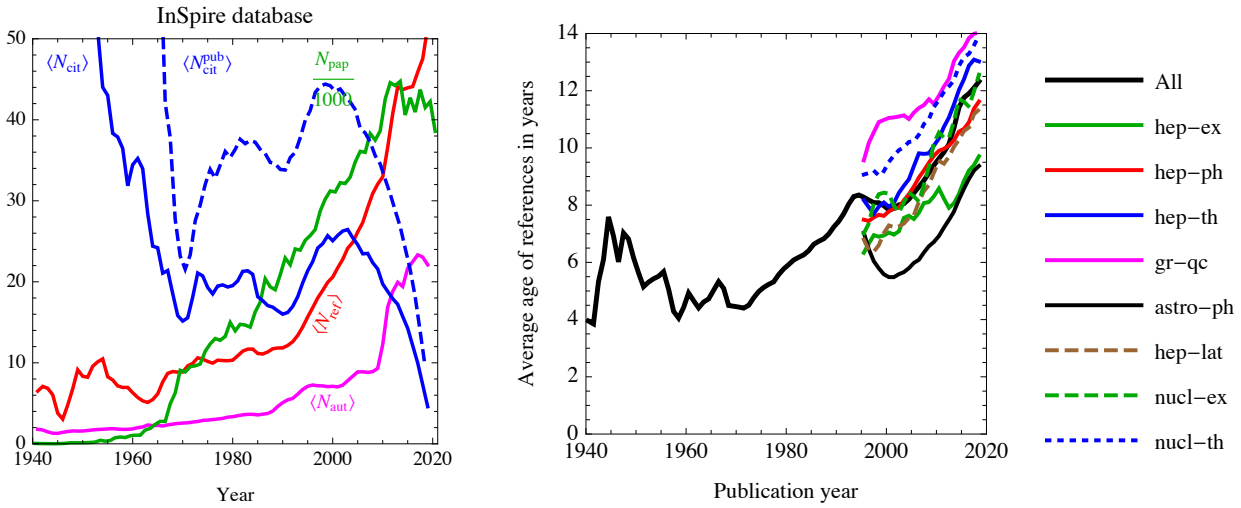

Figure 13: **Left**: *The vertical axis refers to different quantities: number of papers per year (green), average number of references (red), of authors (magenta), of citations (blue), of citations among published papers (blue dashed). Figure 14 shows the same trends within arXiv categories.* **Right**: *average age of references. Results before* $\sim$ 1960 *are affected by the smaller size of the literature, and by the incomplete coverage in the* InSpire *database.*

Even though this should not be interpreted analogously to the Gini coefficient in economy, it is interesting to notice that the typical Gini coefficient of citations is between 0.7 (when looking at single arXiv categories) and 0.8 (when considering the whole database). This means that few papers get most of the citation impact: 4% of papers have more than 100 citations, and they receive half of the total citations; half of the papers have less than 4 citations, and they receive 2% of the total citations (see also Lehmann, Lautrup, and Jackson (2003)).

A.2 Details about the dataset

Any large database contains a small fraction of incomplete/inconsistent information, which may affect any algorithmic study of the data in a variety of ways (del Corso and Romani, 2016). The InSpire database is extremely curated. References are covered with an accuracy at the % level, typically better than big private databases such as Scopus[24] or WebOfScience[25], and comparable to Google Scholar[26].

We obtained the InSpire database in the form of a 'dump' file.[27] InSpire maps papers, authors, collaborations, institutes (affiliations), and journals to record IDs (integer numbers) thereby addressing the problem of name disambiguation (Golosovsky and Solomon, 2017, Mar-

[24] https://www.scopus.com.
[25] https://apps.webofknowledge.com.
[26] https://scholar.google.com.
[27] http://inspirehep.net/dumps/inspire-dump.html.

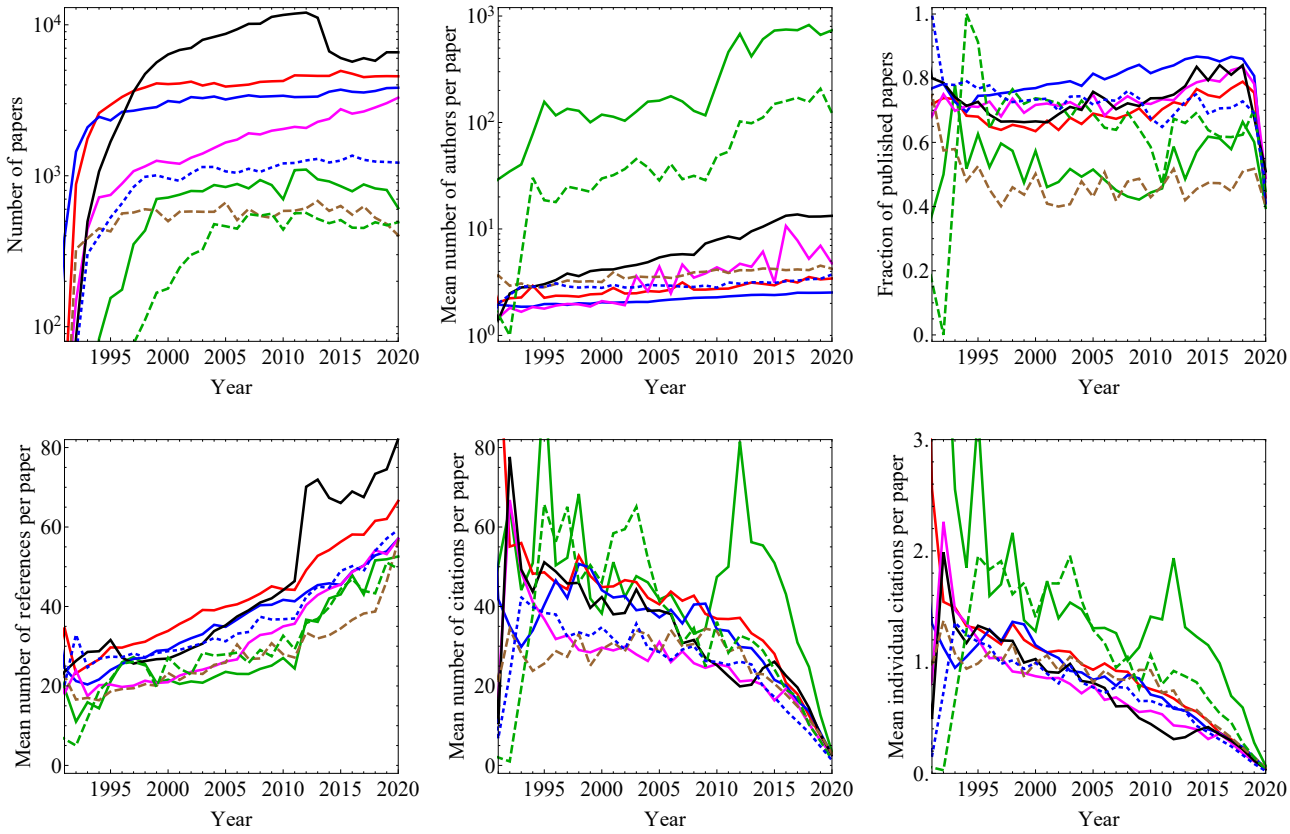

Figure 14: *Trends within the main arXiv representative categories inside* InSpire *(papers included in* InSpire *only).*

tin Montull, 2011).

The extraction of an accurate date for each paper from the database suffers from some uncertainty. There are several available dates: the date when the paper was added to the database, a preprint date, often, but not always, corresponding to the arXiv preprint (when available), a publication date (if the paper has been published on a journal), and an "earliest" date, representing the first available date (not always present). Moreover, month information is typically available only for arXiv papers and some published ones. In general we estimate our uncertainty on the extracted dates at the percent level, in the sense that dates are accurately extracted (at least the year) for about 99% of the papers. Given this uncertainty our sample of papers consists of 1403881 papers from 1230 (de Sacrobosco, 1230) to 31 December 2020.

Another source of uncertainty comes from the author list of each paper. Some papers carry an empty list of authors. This can have different reasons. For instance, only the name of the collaboration is available for experimental papers (mainly conference notes) indexed from the CDS database.[28] All these papers are included in our analysis for what concerns papers, but do not contribute to the metrics for authors. Similar problems extends to institutions and

[28] https://cds.cern.ch.

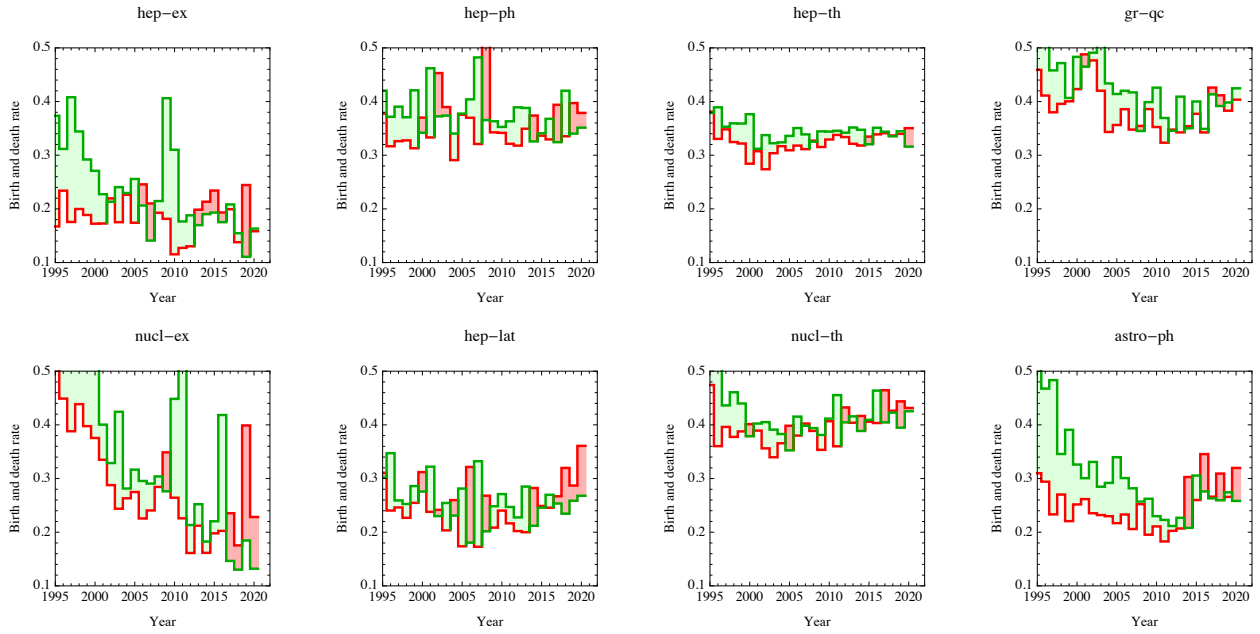

Figure 15: *Percentage of authors appeared or disappeared each year within the arXiv categories fully covered by* InSpire.

journals.

One more source of uncertainty is the extraction of references from papers, needed to produce a citation map. This can be very simple when a bibtex or xml bibliography file is attached to the paper, but can become an extremely complicated task for papers where only a pdf, sometimes produced from a scanned paper, is present. InSpire uses state-of-the-art technology for reference extraction (like Refextract[29] and Grobid[30]), which is mainly automatic, with human supervision only in case of errors and inconsistencies. Despite the advanced technology for reference extraction, not all references are correctly extracted.

There are different kind of problems so that a reference can: simply be missed, be recognised incompletely, be misidentified with another one, be assigned to an inexistent paper ID, point out of the database (in which case it is counted in the number of references, but not indexed), or point to a later paper ("a-causal" reference). All these effects are observed in the database. While some are simple mistakes, or typos in the ID of the paper, the last could be a real effect (below 1%), with the reference appearing in a subsequent version of the paper with no available information on the dates of the different versions. Since "a-causal" citations can generate anomalies in the computation of the rank (typically only when the a-causality is large, i.e. in case of mistakes), we deleted them from our dataset. However, since dates are extracted with an accuracy that is often of one year, we still consider causal all the references to papers with the same date, regardless of the actual papers appearing order.

[29] https://github.com/inspirehep/refextract.
[30] https://github.com/kermitt2/grobid.

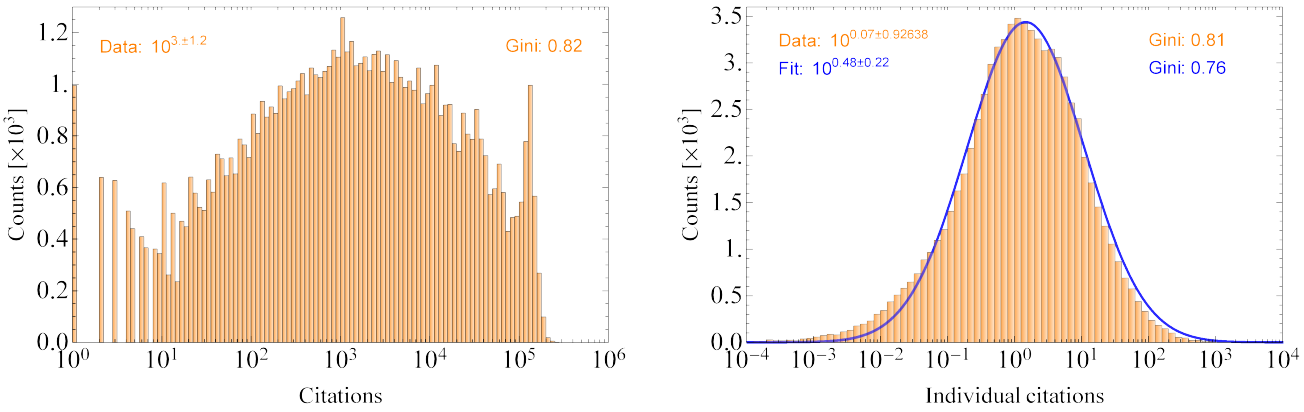

Figure 16: *Distribution of citations (left) and individual citations (right) in the whole* INSPIRE *database. The blue curve in the right panel is the fit to data with a log-normal distribution with the parameters indicated in the caption. Both the distributions give a Gini index close to 0.8.*

In any case, especially because of references pointing outside the database, the number of references indexed for any given record is a better estimate of the actual number of references than the one obtained by summing the indexed ones. Only when computing PaperRanks and AuthorRanks, the number of references is defined equal to the number of indexed references, which is needed to correctly normalize the transition matrix defining the citation graph. Given that INSPIRE is complete only after 1970, this means that references of older papers are over-attributed to those old notable papers that happen to be in INSPIRE.

In summary, the dataset consists of 1403881 papers, 78941 indexed authors, 8321 institutes, 2644 journals, 37172002 references (of which 26333765 indexed). We compute citations directly from references and do not use citation information from the INSPIRE database.

Concerning the information on the arXiv database used in Fig. 12, we imported all records using the arXiv API.[31] All other information on arXiv papers and categories has been obtained from the INSPIRE database. The full list of arXiv categories and subject-classes can be found in the arXiv API reference manual.

All indices discussed in this paper can be computed in one hour on a laptop, apart for the Author Rank, which involves a large, not very sparse matrix: $\sim 8 \cdot 10^4 \times 8 \cdot 10^4$ with about 5×10^8 non-vanishing entries.

References

Aksnes, D. W., Schneider, J. W., and Gunnarsson, M. (2012). Ranking national research systems by citation indicators. A comparative analysis using whole and fractionalised counting methods. *J. Informetrics*, *6*(1), 36–43. (SEMANTIC SCHOLAR) doi: [link]

Bini, D., del Corso, G. M., and Romani, F. (2010). A combined approach for evaluating papers, authors

[31] https://arxiv.org/help/api/user-manual.

and scientific journals. *J. Comput. Appl. Math.*, *234*, 3104-3121. (Semantic Scholar) doi: [link]

Bini, D. A., del Corso, G. M., and Romani, F. (2008). Evaluating scientific products by means of citation-based models: a first analysis and validation. *ETNA, Electron. Trans. Numer. Anal.*, *33*, 1–16. (zbMATH)

Bouyssou, D. and Marchant, T. (2016). Ranking authors using fractional counting of citations: An axiomatic approach. *J. Informetrics*, *10*(1), 183–199. (Semantic Scholar) doi: [link]

Brin, S., Larry Page, R. M., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. *Stanford InfoLab,* (1999). (Semantic Scholar)

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Comput. Netw.*, *30*(1-7), 107-117. (Semantic Scholar) doi: [link]

Carbone, V. (2011). Fractional counting of authorship to quantify scientific research output. *CoRR*, *abs/1106.0114*. (Semantic Scholar)

Chen, P., Xie, H., Maslov, S., and Redner, S. (2007). Finding scientific gems with google's pagerank algorithm. *J. Informetrics*, *1*, 8-15. (Semantic Scholar) doi: [link]

del Corso, G. M. and Romani, F. (2009). Versatile weighting strategies for a citation-based research evaluation model. *Bull. Belg. Math. Soc. - Simon Stevin*, *16*(4), 723–743. (zbMATH)

del Corso, G. M. and Romani, F. (2016). A multiclass approach for ranking graph nodes: models and experiments with incomplete data. *Inf. Sci.*, *329*, 619-637. (Semantic Scholar)

de Sacrobosco, J. (1230). Tractatus de sphaera.

(InSpire)

Ding, Y., Yan, E., Frazho, A. R., and Caverlee, J. (2009). Pagerank for ranking authors in co-citation networks. *JASIST*, *60*(11), 2229–2243. (Semantic Scholar) doi: [link]

Egghe, L. (2008). Mathematical theory of the h- and g-index in case of fractional counting of authorship. *JASIST*, *59*(10), 1608–1616. (Semantic Scholar) doi: [link]

Golosovsky, M. and Solomon, S. (2017). Growing complex network of citations of scientific papers - measurements and modeling. *Phys. Rev. E*, *95*(012324). (Semantic Scholar) doi: [link]

Henneken, E. A. and Kurtz, M. J. (2017). Usage Bibliometrics as a Tool to Measure Research Activity.

(Semantic Scholar)

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proc. Nat. Acad. Sci.*, *46*. (Semantic Scholar) doi: [link]

Holtkamp, A., Mele, S., Simko, T., and Smith, T. (2010). Inspire: Realizing the dream of a global digital library in high-energy physics.

(InSpire)

Hooydonk, G. V. (1997). Fractional counting of multiauthored publications: Consequences for the impact of authors. *JASIS*, *48*(10), 944–945. (Semantic Scholar) doi: [link]

Ivanov, R. and Raae, L. (2010). Inspire: A new scientific information system for hep. *J. Phys. Conf. Ser.*, *219*. (InSpire) doi: [link]

Klem, J. and Iwaszkiewicz, J. (2011). Physicists get inspired: Inspire project and grid applications. *J. Phys. Conf. Ser.*, *331*. (InSpire) doi: [link]

Kurtz, M. J. (2017). Comparing People with Bibliometrics.

(Semantic Scholar)

Kurtz, M. J. and Henneken, E. A. (2017). Measuring metrics - a 40-year longitudinal cross-validation of citations, downloads, and peer review in astrophysics. *Journal of the American Society for Information Science and Technology*, *68*, 695-708. (Semantic Scholar) doi: [link]

Lehmann, S., Lautrup, B., and Jackson, A. (2003). Citation distributions in high-energy physics. *Phys. Rev. E*, *68*. (InSpire) doi: [link]

Leydesdorff, L. and Bornmann, L. (2010). How fractional counting affects the impact factor: Steps towards field-independent classifications of scholarly journals and literature. *CoRR*, *abs/1007.4749*. (Semantic Scholar)

Leydesdorff, L. and Bornmann, L. (2011). How fractional counting of citations affects the impact factor: Normalization in terms of differences in citation potentials among fields of science. *JASIST*, *62*(2), 217–229. (Semantic Scholar) doi: [link]

Leydesdorff, L. and Opthof, T. (2010). Normalization at the field level: fractional counting of citations. *J. Informetrics*, *4*, 644-646. (Semantic Scholar)

Leydesdorff, L. and Park, H. W. (2017). Full and fractional counting in bibliometric networks. *J. Informetrics*, *11*(1), 117–120. (Semantic Scholar) doi: [link]

Leydesdorff, L. and Shin, J. C. (2011). How to evaluate universities in terms of their relative citation impacts: Fractional counting of citations and the normalization of differences among disciplines. *JASIST*, *62*(6), 1146–1155. (Semantic Scholar) doi: [link]

Ma, N., Guan, J., and Zhao, Y. (2008). Bringing pagerank to the citation analysis. *Inf. Process. Manage.*, *44*(2), 800–810. (Semantic Scholar) doi: [link]

Martin Montull, J. (2011). Inspire: Managing metadata in a global digital library for high-energy physics. *Commun. Comput. Info. Sci.*, *240*, 269-274. (InSpire) doi: [link]

Perianes-Rodríguez, A., Waltman, L., and van Eck, N. J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. *J. Informetrics*, *10*(4), 1178–1195. (Semantic Scholar) doi: [link]

Pinski, G. and Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Inf. Process. Manage.*, *12*(5), 297–312. (Semantic Scholar) doi: [link]

Radicchi, F., Fortunato, S., Markines, B., and Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Phys. Rev. E*, *80*, 056103. (Semantic Scholar) doi: [link]

Rajaraman, A. and Ullman, J. D. (2009). Mining of massive datasets. *Cambridge University Press*. (Semantic Scholar) doi: [link]

Rousseau, R. (2014). A note on the interpolated or real-valued h-index with a generalization for fractional counting. *Aslib J. Inf. Manag.*, *66*(1), 2–12. (Semantic Scholar) doi: [link]

Sinatra, R., Deville, P., Szell, M., Wang, D., and Barabási, A.-L. (2015). A century of physics. *Nature Physics*, *11*(10), 791–796. (Semantic Scholar) doi: [link]

Tanase, R. and Radu, R. (2009). *The Mathematica of Web Search*.

Thelwall, M. and Wilson, P. (2014). Regression for citation data: An evaluation of different methods. *J. Informetrics*, *8*(4), 963 - 971. (Semantic Scholar) doi: [link]

Walker, D., Xie, H., Yan, K.-K., and Maslov, S. (2007). Ranking scientific publications using a simple model of network traffic. *J. Stat. Mech.*, *0706*. (Semantic Scholar) doi: [link]

Wang, D., Song, C., and Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127–132. (Semantic Scholar) doi: [link]

West, J. D., Jensen, M. C., Dandrea, R. J., Gordon, G. J., and Bergstrom, C. T. (2013). Author-level eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. *JASIST*, *64*(4), 787–801. (Semantic Scholar) doi: [link]

Zhou, D., Orshanskiy, S. A., Zha, H., and Giles, C. L. (2007). Co-ranking authors and documents in a heterogeneous network. *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007)*, 739–744. (Semantic Scholar) doi: [link]