nature REVIEWS PHYSICS

SUPPLEMENTARY INFORMATION

In format as provided by the authors

Taking census of physics

Federico Battiston, Federico Musciotto, Dashun Wang, Albert-László Barabási, Michael Szell and Roberta Sinatra

https://doi.org/10.1038/s42254-018-0005-3

Taking Census of Physics

Supplementary Information

Federico Battiston[1,*], Federico Musciotto[1,*], Dashun Wang[2,3], Albert-László Barabási[1,4,5], Michael Szell[1,4,6], and Roberta Sinatra[1,7,4,8,+]

[1]Department of Network and Data Science, Central European University, Budapest, 1051, Hungary

[2]Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA

[3]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208, USA

[4]Network Science Institute, Northeastern University, Boston, MA 02115, USA

[5]Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA

[6]MTA KRTK Agglomeration and Social Networks Lendulet Research Group, Centre for Economic and Regional Studies, Hungarian Academy of Sciences, Budapest, 1094, Hungary

[7]Department of Mathematics, Central European University, Budapest, 1051, Hungary

[8]ISI Foundation, Torino, 10126, Italy

[*]These authors contributed equally to this work.

[+]Correspondence should be addressed to: robertasinatra@gmail.com

S1 Defining physics publications in non-physics journals

We identify physics publications in journals which are not explicitly labelled as physics journals by means of a method first used in Refs.[1,2] Such method allows us to reconstruct a community in a network when only a small fraction of nodes are explicitly labelled as belonging to the community. In our case, the hypothesis is that physics papers can be found not only in conventional physics journals (core physics papers) but also in other venues (interdisciplinary physics papers). It is possible to identify such interdisciplinary papers if they have a significant number of connections (references or citations) to conventional physics venues, applying the conceptual framework of label propagation algorithms.[3] In Ref.[1] the algorithm was first applied to an old version of the Web of Science (WoS), encoding information about scientific publications until 2012 and based on an old database structure. Here we reapply the method on an updated version of WoS purchased from Clarivate Analytics, encoding information about publications until 2017, and using a new database structure, with a different identification system for papers among other things. We obtain a new physics dataset of papers, which we want to further characterise by identifying the physics subfields they belong to. For this reason, all papers in the dataset except those of journals published by the American Physical Society (APS) journals, are then considered to be assigned a given subfield and be part of our physics communities analysis. The label propagation method for the identification of subfields, illustrated in detail in Section S3, is a modified implementation of the algorithm to identify physics papers presented next.

The algorithm to construct the physics dataset works in the following way. Let us consider a directed network with N nodes, for instance the citation network described by the WoS dataset, where nodes are scientific publications, and a direct link between publication i and publication j exists if paper i cites paper j. Each node i has an in-degree k_{IN} (number of citations) and an out-degree k_{OUT} (number of references). Nodes with $k_{IN} = 0$ and $k_{OUT} = 0$ are publications without references and citations and are isolated nodes in the network. Additionally, in our case each node i is characterised by a variable t_i corresponding to the time of publication of the article. The method is based on an iterative process where at each step s the N nodes are assigned to three sets: the core set C_s, the tangent set T_s and the external set E_s. The core set C_s includes the nodes that are considered to be part of the target community at a given time step s by the algorithm. In our case, at the step $s = 0$, C_0 includes all articles published in physics journals. The purpose of this initial core set is to act as a seed to detect other nodes that are part of the community, even if initially they are not classified as such, and that will be iteratively included in C_s at subsequent steps $s > 0$. The second set is the tangent set T_s, and contains all the nodes outside the core set C_s that have at least one (ingoing or outgoing) connection to a node within C_s. The third set is the external set E_s, and corresponds to all nodes outside the core set C_s that share no connection with nodes within C_s, and therefore have no chance to be included into the core at the subsequent step $s+1$. By definition we have $C_s \cup T_s \cup E_s = N$ and $C_s \cap T_s \cap E_s = \emptyset$.

The basic idea of the algorithm is to iteratively extend the target community C_s into C_{s+1} by adding candidate nodes from T_s that are statistically expected to be part of the community based on their connections. In our case this corresponds to identifying as physics all scientific papers which are not published in physics journals, but whose patterns of references and citations are indistinguishable from those published in the traditional physics venues. The purpose of the tangent set T_s is to contain all candidate nodes, i.e. nodes that might subsequently be added to the target community C_s at step s after inspection of their incoming and outgoing links. To do so, at each step s and for each node i we compute two variables: $r_{i,s}^{IN}$ and $r_{i,s}^{OUT}$. These variables quantify the expectation of a particular node to be part of the target community C_s based on its incoming citations and outgoing references.

Let us focus first on incoming citations, evaluated through $r_{i,s}^{IN}$, where

$$r_{i,s}^{IN} = \frac{k_{i,s}^{IN,\odot}}{\hat{k}_{i,s}^{IN,\odot}}. \tag{1}$$

Here $k_{i,s}^{IN,\odot}$ corresponds to the number of incoming links (citations) to node i originating from nodes in the core C_s. $\hat{k}_{i,s}^{IN,\odot}$, instead, accounts for the expected number of incoming links from the core in a null model where the real number of incoming and outgoing links of each node (citations and references of each paper) in the network is fixed. This last constraint corresponds to consider the directed configuration model ensemble of the original citation network, meaning that we can write

$$\hat{k}_{i,s}^{IN,\odot} = k_i^{IN} \frac{\sum_{j \in C_s} k_j^{OUT}}{\sum_{j \in N} k_j^{OUT}} \tag{2}$$

where k_i^{IN} denotes the total number of incoming links to node i, and the remaining term corresponds to the probability for a link to originate from C_s. As an article i can receive a citation from another paper j only if the latter is more recent, i.e. $t_j > t_i$, we eventually set

$$\hat{k}_{i,s}^{IN,\odot} = k_i^{IN} \frac{\sum_{j \in C_s | t_j > t_i} k_j^{OUT}}{\sum_{j \in N | t_j > t_i} k_j^{OUT}}. \tag{3}$$

Similarly, the share of outgoing references are evaluated through $r_{i,s}^{OUT}$, where

$$r_{i,s}^{OUT} = \frac{k_{i,s}^{OUT,\odot}}{\hat{k}_{i,s}^{OUT,\odot}}, \tag{4}$$

and

$$\hat{k}_{i,s}^{OUT,\odot} = k_i^{OUT} \frac{\sum_{j \in C_s | t_j < t_i} k_j^{IN}}{\sum_{j \in N | t_j < t_i} k_j^{IN}}. \tag{5}$$

A value $r_{i,s}^{IN} > 1$ ($r_{i,s}^{OUT}$>1) corresponds to a node that is more likely to reference (be cited from) nodes from the core than what would be expected at random. At each step s of the process, we use the variables $r_{i,s}^{IN}$ and $r_{i,s}^{OUT}$ associated to nodes in T_s to produce the updated core set C_{s+1}. First we add all nodes in C_s to C_{s+1}. Then, for each node $i \in T_s$, we add i to C_{s+1} if we have

$$r_{i,s}^{IN} > \tau^{IN} \tag{6}$$

or

$$r_{i,s}^{OUT} > \tau^{OUT}. \tag{7}$$

The thresholds τ^{IN} and τ^{OUT} are fixed based on a parameter p such that the thresholds τ^{IN} and τ^{OUT} correspond respectively to the $p-th$ percentile of the distribution of $r_{i,0}^{IN}$ and $r_{i,0}^{OUT}$ values for nodes within the initial core set C_0. Once nodes $i \in T_s$ satisfying the conditions of Eq.6 or Eq.7 are added to the core set C_{s+1}, both sets T_s and E_s can be updated to T_{s+1} and E_{s+1} from C_{s+1}. The process stops when C_s has converged, i.e. when no nodes from T_s can be added to the core set C_s. Note that while the thresholds τ^{IN} and τ^{OUT} remain constant during the whole process, the values $r_{i,s}^{IN}$ and $r_{i,s}^{OUT}$ associated to each node i will change at each iteration, given the fact that new nodes will incorporate the set C_s at each iteration step s. As shown in Ref.,[2] in the case of physics publication in the WoS dataset the algorithm was run iteratively for 10 steps, showing fast convergence.

The parameter p is a tolerance parameter, in the sense that it defines the minimal attraction needed for a node to be incorporated in the growing core. As described in Refs.,[1,2] it is possible to set the value of p by validating the algorithm on all publications of interdisciplinary journals for which a subset is labelled explicitly as physics. In our case, we use *Science* (1995-2013) and *PNAS* (1915-2013) for this validation. The best trade-off between true positive (92.3%) and true negative rates (99.6%) was found for $p = 10$. By running the algorithm on the new version of the WoS dataset comprised of ~54 million papers, with an initial core of ~3.2 million articles published in 294 physics journals, the list of journals being extracted by combining information from Wikipedia, Scopus and Scimago, we identified ~4.5 million physics publications in non-physics journals. In Table S1 we report the ten non-physics journals with the highest number of physics publications (number of papers in brackets), and the ten non-physics journals with the highest share of physics publications (percentages in brackets). We note the presence of interdiscisciplinary journals, such as *Nature*, and several materials and chemistry journals. In the main text we focused our analysis on the period 1985-2015, restricting our dataset to a total of ~5.6 million publications.

Rank	**Journal** (number of papers)	**Journal** (percentage of papers)
1	Rev. Sci. Instrum. (41,006)	J. Space Weather Space Clim. (98.3%)
2	Thin Solid Films (38,316)	Quantum Inform. Comput. (97.7%)
3	Surf. Sci. (34,461)	J. Hyberbolic Differ. Equ. (93.2%)
4	Nature (33,073)	Adv. Quantum Chem. (92.8%)
5	J. Alloy. Compd. (31,319)	2D Mater. (92.7%)
6	J. Phys. Chem. (29,920)	Thin Solid Films (92.2%)
7	J. Am. Chem. Soc. (27,192)	J. Laser Micro Nanoeng. (92.2%)
8	Macromolecules (26,377)	Surf. Sci. Rep. (91.9%)
9	Electron. Lett. (25,593)	IEEE Trans. Nanotechnol. (91.4%)
10	J. Electrochem. Soc. (24,806)	Symmetry Integr. Geom. (90.1%)

Table S1. Non-physics journals with most physics publications and highest percentage of physics publications identified by means of label propagation.

S2 Identifying physics subfields from PACS codes

Despite the WoS dataset provides a classification of core physics publications into different subfields (see Section S3), such classification is not detailed enough to our scope and, most importantly, it fails to associate a subfield to publications not in physics journals. For such a reason, in our work we associated publications to different subfields according to the Physics and Astronomy Classification Scheme (PACS) by the American Institute of Physics,[4] a hierarchical classification used by several journals, including papers of the Physical Review Series published by the American Physical Society between 1977 and 2015. The classification uses four digits and an extra identifier. The 1-digit identifies 10 different physics subfields, namely: General (0), The Physics of Elementary Particles and Fields (shortened as HEP, 1), Nuclear Physics (2), Atomic and Molecular Physics (AMO, 3), Electromagnetism, Optics, Acoustics, Heat Transfer, Classical Mechanics, and Fluid Dynamics (Classical, 4), Physics of Gases, Plasmas, and Electric Discharges (Plasma, 5), Condensed Matter: Structural, Mechanical and Thermal Properties (6), Condensed Matter: Electronic Structure, Electrical, Magnetic, and Optical Properties (7), Interdisciplinary Physics and Related Areas of Science and Technology (Interdisc, 8), Geophysics, Astronomy, and Astrophysics (Astro, 9). We merged PACS 6 and 7 into a unique category named CondMat, in order to match other common physics classifications, such as that found for the arXiv (see Section S3). We stress that the term interdisciplinary physics, assigned in Ref.[1] to describe physics publications in non-physics journals, is not linked to the PACS 8 of the PACS scheme. In the following, as well as in the main text, the term Interdisciplinary physics is reserved to identify publications and authors working in this precise subfield of physics, differently from Ref.[1] PACS can be found in the APS dataset, available from the APS upon request,[5] encoding information about all publications appeared in the journals of the American Physical Society until 2015. Although PACS appeared in 1977, only a small fraction of the papers were assigned one until they were enforced in 1985. For this reason, we focused our analysis on the years 1985-2015, for which our dataset has 435,722 papers with at least one PACS. 5,616 more papers have assigned a PACS but were published before 1985. More in detail, between 1985-2015 we have 265,549 papers with exactly one 1-digit PACS, 138,176 with two PACS, 29,806 with three PACS, 2,160 with PACS and 31 with five PACS.

In Fig. S1 we report the distribution of the 9 physics subfields for six well-established journals published by the APS, namely the general purpose *Physical Review Letters* and the specialised venues *Physical Review A - E*. *Physical Review B* (covering condensed matter and materials physics) and *Physical Review C* (covering nuclear physics) indeed predominantly publish papers belonging to a single subfield, respectively Condensed Matter and Nuclear Physics. Conversely *Physical Review A* (covering atomic, molecular, and optical physics and quantum information), *Physical Review D* (covering particles, fields, gravitation, and cosmology) and *Physical Review E* (covering statistical, nonlinear, biological, and soft matter physics) publish across a greater mixture of subfields. As expected, *Physical Review Letters*, the APS flagship journal, publishes across all different domains, even though with different frequency.

Similarly to the identification of physics papers in non-physics venues, we use the papers published in the APS journals as the initial seed to assign subfields to other physics publications by means of label propagation (see Section S3 for details.). In such a way, we obtain a data-driven subfield classification of physics papers in the WoS dataset.

In Fig. S2a we report the proportions of APS papers belonging to a given subfield, and compare it to that of our newly created dataset. In Fig. S2b we report the distribution of the number of subfield per paper in the APS between 1985 and 2015, as well as the fraction of

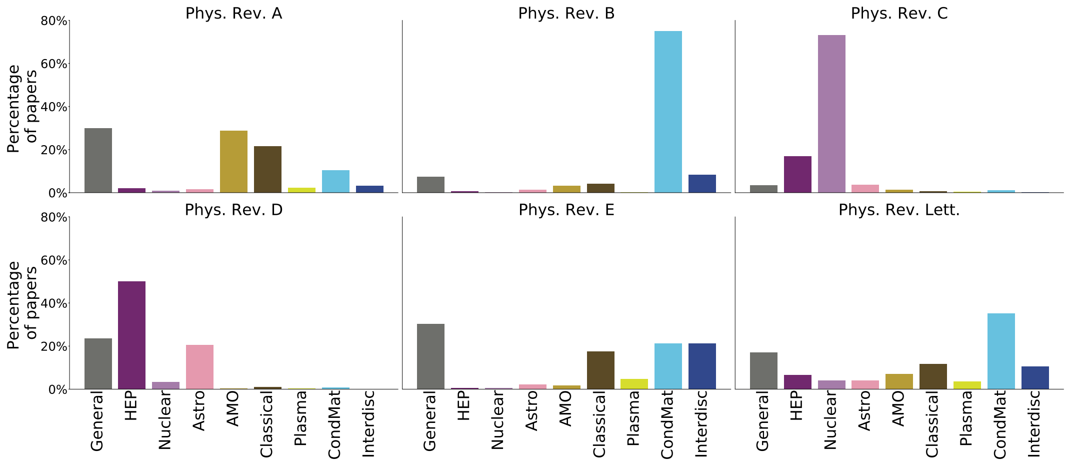

Figure S1. Subfield distribution for papers published in APS journals. Different APS journals show different publication patterns across subfields. *Physical Review B* covers predominantly CondMat, and *Physical Review C* is similarly focused on Nuclear. In contrast, *Physical Review A*, *Physical Review D* and *Physical Review E* do not cover a single, predominant subfield. *Physical Review Letters* is the most balanced journals of the APS publishing across all subfields.

number of papers per subfield over the years (Fig. S2c).

S3 Assigning Physics subfields to Web of Science publications

We propagate physics subfields to physics publications in the WoS dataset based on relevant patterns of references and citations to the specific subfield(s), adapting the method described in the first section of this SI. For each subfield we have a different initial core set C_0^α, corresponding to all publications in the APS publications between 1985 and 2015 associated to a given subfield α. First, we matched the papers of the APS dataset into the Web of Science dataset, either via exact doi matching, or, for when the doi is not available, by using the Levenshtein distance to compute title similarity. In this second case the match was accepted if there was at least 90% string similarity between the titles of two papers in the datasets, and the second best match had a string similarity at least 5 times worse. In this way we were able to match 90% of all the papers manually assigned to a subfield between 1985 and 2015.

In the original implementation, the thresholds τ^{IN} and τ^{OUT} were set by evaluating the performance of the algorithm on the "ground truth" of physics papers published in interdisciplinary journals such as *Science* and *PNAS*. However, such a validation is not possible at the subfield level. Hence, for label propagation at the subfield level we slightly modified the original implementation. We observe that the algorithm may propagate subfields both to papers within and out of the original APS core, which is made of papers that already have a PACS code. For such a reason, for each subfield α we selected the threshold τ^{α} so that after 10 iterations the number of papers of each subfield cannot grow more than 10% within the original APS dataset. For simplicity, we chose $\tau^{IN,\alpha} = \tau^{OUT,\alpha}$. Afterwards, we performed label propagation for each subfield α independently. We obtained a total of 1,137,670 papers in WoS published between 1985 and 2015 and classified within one of the subfields of Physics. We note that also some papers outside the considered time-span were assigned a subfield, but we focused our analysis on the period $1985 - 2015$ to be consistent with the years when PACS were systematically used in publications by the APS. As already mentioned, PACS corresponding to the two categories associated to Condensed Matter were merged into the same subfield.

We compare the classification of papers obtained through label propagation with that of the original APS dataset by measuring the fraction of subfields in the original and the propagated datasets (Figure S2a). The two datasets have a similar subfield distribution with a cosine similarity of 0.99. Differences in the two datasets are likely to indicate an under- or over- representation of some areas of physics in the *Physical Review* series compared to the overall physics world. In Fig. S2b we report the distribution of the number of subfields per paper in the two datasets. Papers in the reconstructed physics dataset tend to be slightly more specialised (70% of the papers are assigned to a single subfield) than those in the APS dataset (61%). However, overall the two distributions are quite similar. Finally, in Figs. S2c,d we show the evolution of the fraction of papers of different subfields in the APS dataset and in our reconstructed dataset from 1985 to 2015. It is evident how the two datasets have very similar temporal patterns during the period under investigation.

Validation: To test the robustness of our findings, we validated our data-driven classification of papers across subfields. As already mentioned, PACS codes were systematically introduced in publications in the APS journals 1985. As our method classifies papers into subfields according to patterns of references and citations only, our algorithm naturally assigns subfields also to publications in the APS journals before 1985, provided that they are significantly connected to the corresponding core papers for the subfield(s). Five of the previously six analysed APS

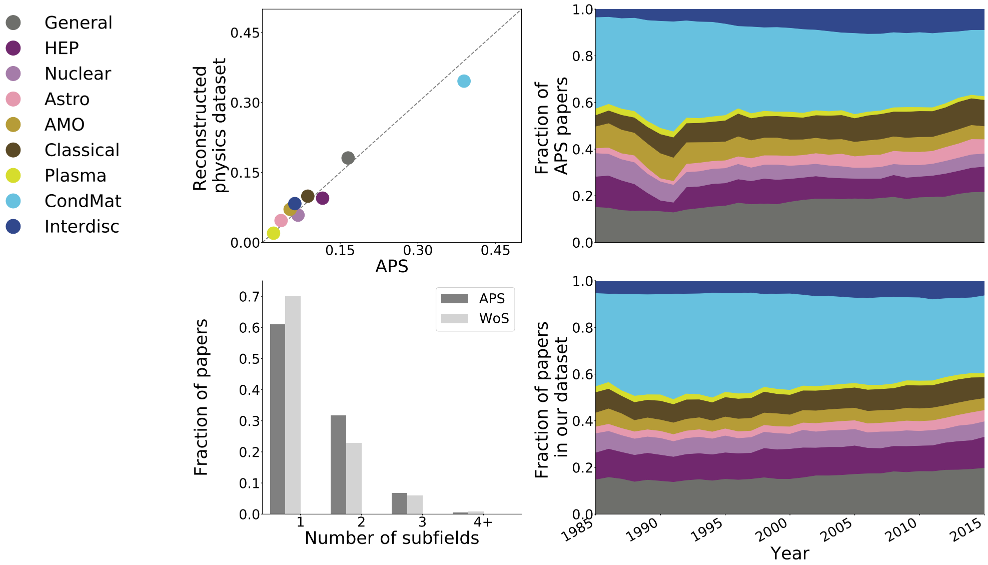

Figure S2. Comparison between the APS dataset and the reconstructed physics dataset. a Scatterplot of the fraction of subfields appearing in papers of the APS dataset and in the reconstructed physics dataset. **b** Distribution of number of subfields per paper in the two datasets. **c**, **d** Temporal evolution of the fraction of subfields between 1985 and 2015 for the two datasets. Colours for subfields are consistent with those used in the main text.

journals (with the exception of *Physical Review E*) were born before 1985. In Fig. S3 we test the robustness of the subfield distributions in the journals as a way to assess the effectiveness of our data-driven method to classify physics papers across subfields by comparing the distribution of the subfield manually assigned between 1985 and 2015 in *Physical Review. A, B, C, D*, and *Physical Review Letters*, with that obtained by means of label propagation for papers published before 1985 in the same journals. The two distributions are highly correlated for all journals, with cosine similarities ranging from 0.88 to 0.99 .

We also tested the robustness of our subfield categorisation by comparing it to additional sources providing alternative physics classifications, namely the physics classification provided by *(i)* the WoS dataset (for core physics papers only), (ii) the arXiv repository, that collects electronic preprints of papers related to physics topics. The cosine similarity between the fraction of papers in our dataset and in the two alternative datasets is quite high, respectively *(i)* 0.86 for WoS, *(ii)* 0.74 the arXiv. The scatterplots between our reconstructed physics dataset and the other databases are shown respectively in Figs. S4a,b. Achieving a perfect mapping between

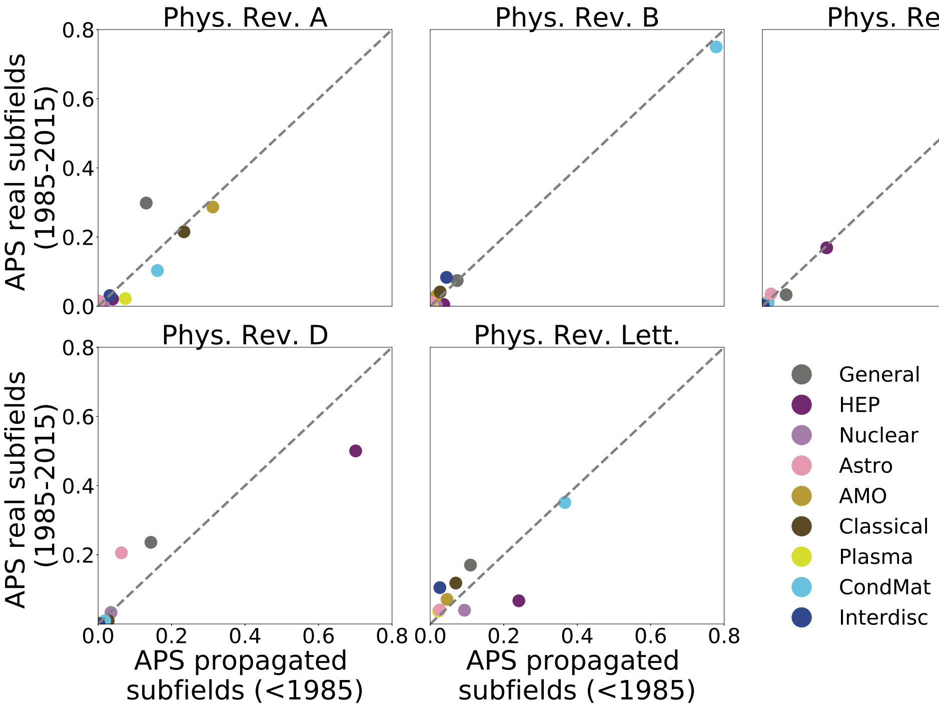

Figure S3. Testing propagated subfields in APS journals before 1985. Scatterplot between the subfield distribution of the papers published in the APS journals after 1985, and the propagated subfield distribution for papers published before 1985 in the same journals. The cosine similarities between the distribution of papers before and after 1985 are *(i)* 0.91 for *Physical Review A*, *(ii)* 0.99 for *Physical Review B*, *(iii)* 0.99 for *Physical Review C*, *(iv)* 0.93 for *Physical Review D* and *(v)* 0.88 for *Physical Review Letters*.

the scheme of arXiv and WoS into the PACS scheme is not possible. As an example, the *nonlin* category in the arXiv dataset, that we eventually mapped into the General physics subfield, actually contains papers of at least an additional subfield, i.e. Interdisc. For the same reason some of the subfields obtained from the PACS scheme do not have a direct counterpart in the other two datasets. However, small changes in the mapping scheme do not modify significantly the output of our validation. For example, merging Classical and Plasma in our dataset to better match the “Fluids & Plasma Physics” category in Web of Science changed the corresponding cosine similarity by less than 1%. We report the full mappings in Table S2. We mention that in the past discrepancies between the arXiv categories assigned manually to papers and data-driven cluster of papers based on co-citations across subfields have been observed.[6]

Another factor that may affect the matching is the presence of specific biases for each of

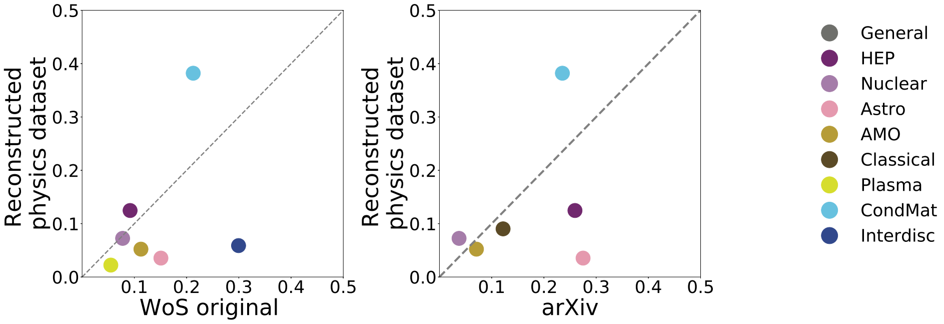

Figure S4. Comparison between the distribution of subfields in our reconstructed physics dataset with the WoS and the arXiv physics categories. Correlation between distributions is high, with values of cosine similarity respectively equal to **a** 0.86, and **b** 0.74. Colours for subfields are consistent with those used in the main text.

these datasets, which are captured by comparing it with our new data-driven reconstructed physics dataset. For instance, the arXiv, first created as a repository for people working on High Energy Physics, shows a disproportionally high number of HEP and Astro publications. This comes as no surprise since the initial scope of the arXiv was to diffuse scientific results in HEP, and the repository has been largely used by such community. In Table S3, we report the five non-APS journals with most papers assigned to each subfield by means of label propagation (number of papers in brackets).

We note that the Astrophysics literature seems to be relatively disconnected to its APS core, compared to results for the other subfields. As an example, we focus on a well established specialised journal in the area, the *Astrophysical Journal*, for which WoS indexes 98,482 papers, only 2,330 of which are labeled. This is because, out of the 3,724,542 outgoing references from papers published in the *Astrophysical Journal*, only 0.6% are directed towards the Astro core. Similarly, out of the 4,896,146 incoming citations towards papers published in the *Astrophysical Journal*, only 1.4% come from the Astro core. As a reference, we compare these numbers with those of *Solid State Communications*, a specialised journal in the area of Condensed Matter, for which our method assign a subfield to 16,274 out of 35,781 papers. In such case, of the 489,625 references and 635,466 citations of the journal, 5.3% and 4.8% link to the CondMat core. These numbers are roughly fives times higher than those for the *Astrophysical Journal*. As a consequence of this disconnection, it is possible that our method it is underestimating the number of (possibly specialised) scientists working in Astrophysics. For both journals the

WoS category	Subfield	arXiv category
/	General	/
Fields	HEP	hep-ex, hep-lat, hep-ph, hep-th, math-ph
Nuclear Physics	Nuclear	nucl-ex, nucl-th
Astrophysics	Astro	astro-ph, gr-qc
Atomic, Molecular & Chemical Physics	AMO	quant-ph
/	Classical	physics, nlin
Fluids & Plasmas Physics	Plasma	/
Condensed Matter Physics	CondMat	cond-mat
Multidisciplinary Physics	Interdisc	/

Table S2. Mapping of physics categories from arXiv categories and WoS physics categories into physics subfields.

fraction of citations (references) coming from (going towards) the cores associated to the other subfields is negligible.

At last, in Fig. S5 we report the publication profile across subfields for three leading interdisciplinary journals. Unsurprisingly, most subfields are represented in all three venues. We note that the proportions of the different subfields is similar to that of the publication of the APS flagship journal, *Physical Review Letters*.

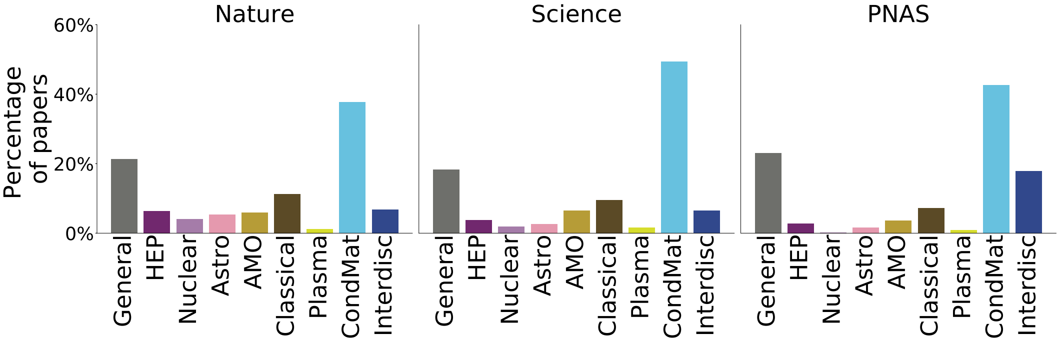

Figure S5. Shares of subfields for publications in *Nature*, *Science* and *PNAS*. All three interdisciplinary journals publish across all subfields of physics.

Rank	General	HEP	Nuclear
1	Phys. Lett. A (9,027)	Nucl. Phys. B (14,524)	Nucl. Phys. A (16,680)
2	J. Phys. A-Math. Gen. (6,029)	J. High Energy Phys. (10,860)	Phys. Lett. B (9,017)
3	Physica A (4,863)	Nucl. Phys. A (7,109)	J. Phys. G-Nucl. Part. Phys. (4,916)
4	Class. Quantum Gravity (4,299)	Prog. Theor. Phys. (5,280)	Nucl. Instrum. Methods Phys. A (4,238)
5	J. Math. Phys. (3,366)	Eur. Phys. J. C (4,462)	Eur. Phys. J. A (2,848)

Rank	Astro	AMO	Classical
1	Phys. Lett. B (3,559)	J. Phys. B (7,005)	Opt. Commun. (4,179)
2	J. Cosmol. Astropart. Phys. (2,370)	J. Chem. Phys. (2,314)	Phys. Lett. A (3,955)
3	Astrophys. J. (2,330)	Nucl. Instrum. Methods Phys. B (1,483)	Opt. Lett. (2,492)
4	Class. Quantum Gravity (2,099)	Phys. Lett. A (1,460)	J. Opt. Soc. Am. B (2,319)
5	Nucl. Phys. B (1,219)	Phys. Scr (1,089)	J. Appl. Phys. (2,123)

Rank	Plasma	CondMat	Interdisc
1	Phys. Plasmas (2,749)	J. Appl. Phys. (23,364)	Physica A (1,977)
2	Phys. Fluids (1,088)	Appl. Phys. Lett. (20,196)	J. Phys. A-Math. Gen. (1,923)
3	Rev. Sci. Instrum (908)	Physica C (19686)	J. Chem. Phys. (1,636)
4	Nucl. Instrum. Methods Phys. A (873)	Solid State Commun. (16,274)	Phys. Lett. A (1,419)
5	Plasma Phys. Control. Fusion (823)	Physica B (15,247)	J. Appl. Phys. (1,297)

Table S3. Non-APS journals with most publications with propagated subfields.

S4 Assigning physicists to subfield(s)

While papers are directly associated to subfields through label propagation, we still need to assign physicists to their correct research area. Some physicists, in particular those extremely productive, are likely to appear over a whole career as the authors of papers belonging to multiple subfields, though some of these might not be significant. As a consequence, when assigning the authors to the different subfields, we applied a statistical filter in order to assign only the subfield(s) on which their engagement is significant. In particular, we consider a physicist as significantly working in a subfield only if her share of publications in it, compared to her production across all subfields, is greater than that of the average scientist. Let us consider the bipartite weighted network $W = \{w_{i\alpha}\}$, where $w_{i\alpha}$ is an integer corresponding to the number of publications of author i in subfield α. The previous condition can hence by formalised as

$$RCA = \frac{\frac{w_{i\alpha}}{\sum_{\alpha'} w_{i\alpha'}}}{\frac{\sum_{i'} w_{i'\alpha}}{\sum_{i'\alpha'} w_{i'\alpha'}}.} > 1. \tag{8}$$

This filter, known as the Revealed Comparative Advantage (RCA) index, was introduced in 1965 in Ref.[7] and has been used previously to filter bipartite networks, as in Ref.[8] Differently from other alternatives, it guarantees that each author is active on at least one field. We limit our analysis to authors with at least $N = 5$ publications in our reconstructed physics dataset, in order to drop all the authors whose contribution to physics is marginal. This set covers 135,877 authors.

The average distribution $w_{i\alpha}$ of subfields per author is shown in Fig. S6a. In Fig. S6b we show the average fraction of papers in each subfield for authors statistically validated in a given area. This plot is similar to that of Fig. **??**c of the main text, but reports more fine-grained information about the involvement of physicists in the subfields to which they are assigned. As shown, the share of publication in the subfield of belonging is the highest for authors in Cond Mat, HEP and Nuclear. Last, in Fig. S6c we report the average career length measured in years, of physicists starting publishing in a given year. As expected, the earlier the starting year, the longer the average time span between the first and last publications of a physicist.

Validation: To test the robustness of our subfield categorisation at the author level, we compared the numbers of authors working in each subfield with the number of APS members registered across APS Divisions.[9] In Fig. S7 we report the scatterplot between the two datasets,

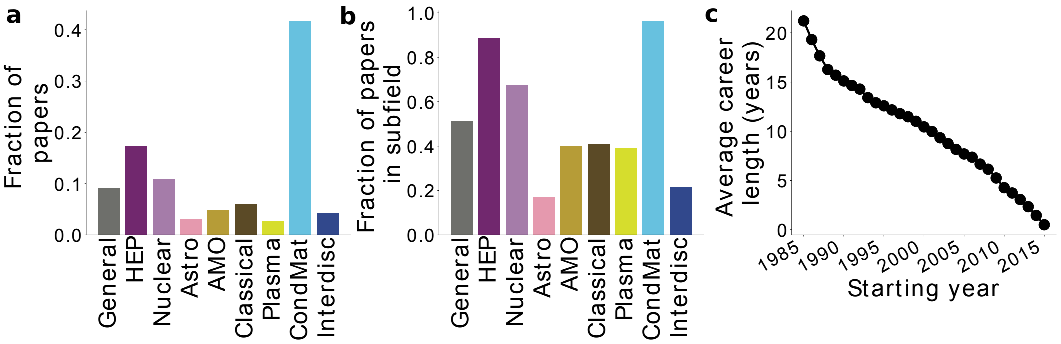

Figure S6. Basic features of authors in our reconstructed physics dataset. a Average publication shares across subfields of a physicist. **b** For authors validated in a subfield, average fraction of publications in that subfield. **c** Average career length measured in years as a function of the starting year of a career.

with a cosine similarity of 0.98. The full mappings between the APS Divisions and our subfield scheme is reported in Table S4.

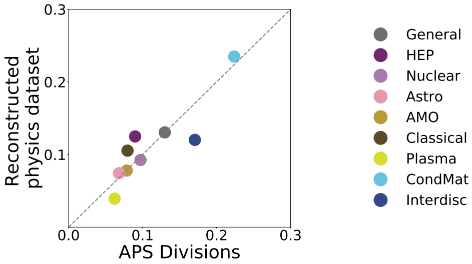

Figure S7. Comparison of the fraction of physicists associated to the different subfields and the members of the APS Divisions. Correlation between the two distributions is high, with a cosine similarity of 0.98. Colours for subfields are consistent with those used in the main text.

S5 Author disambiguation

A common problem in the analysis of scientific careers is that of author disambiguation.[10] Our census of physics is based on merging paper information on subfield and author information on publications provided by the WoS. Our analysis has been undertaken on the latest available version of WoS which, differently from the previous one, has a built-in author disambiguation, where authors are not classified by a name but by a specific author ID. A single author ID is

Subfield	APS Divisions
General	Computational Physics, Quantum Information, Gravitation
HEP	Particles & Fields
Nuclear	Nuclear Physics, Physics of Beams
Astro	Astrophysics
AMO	Atomic, Molecular & Optical
Classical	Fluid Dynamics
Plasma	Plasma Physics
CondMat	Condensed Matter Physics, Laser Science, Polymer Physics
Interdisc	Biological Physics, Materials Physics, Chemical Physics

Table S4. Mapping of physics categories from the APS Divisions into the physics subfield scheme.

associated to a unique author, and can be associated to several author names when the publications authored by the same individual report slightly different name formats. Similarly, two homonyms, but distinct individuals with the same author name are associated to different author IDs. Nevertheless, we are aware that a perfect disambiguation is a goal which is impossible to achieve. For such a reason, we decided to test the robustness of our results by replicating the analysis reported in the main text after excluding a subset of authors with names which are known to be particularly hard to disambiguate. In particular, we focused on the most common 100 Chinese and 200 Korean names,[11,12] which correspond to 504,538 distinct author IDs in the WoS dataset, 15,982 of which are present also in our subset of physicists. Overall, results were shown to be extremely robust to the elimination of such authors. As an example, we report in Fig. S8 the starting point of our analysis, i.e. the authors distribution across subfields. The cosine similarity between the distribution across subfields of the full set and the reduced set of physicists, without authors difficult to disambiguate, is 0.99.

It is worth to mention that highly curated data-repositories with very good author disambiguation is available for some subfields. For instance, the well-known HEP-INSPIRE dataset has an extremely valid author disambiguation, especially needed for fields where most publica-

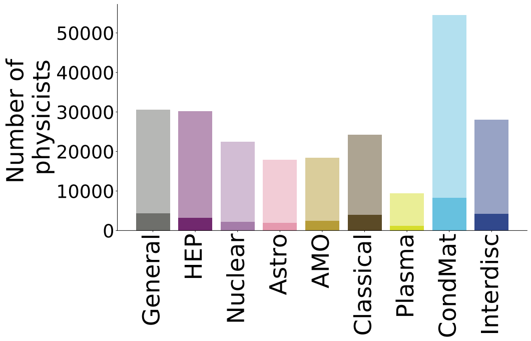

Figure S8. Testing author disambiguation. Number of authors working in each subfield: plain color (reduced set of 15,982 authors difficult to disambiguate), faded color (all other physicists). The cosine similarity between the distribution across subfields of the full set of physicists, and the set without authors hard to disambiguate, is 0.99. Colours for subfields are consistent with those used in the main text.

tions are done by large collaborations. However, it is difficult to map the HEP-INSPIRE author disambiguation into the built-in WoS author disambiguation. On top of this, we believe that such merge would not add validity to our analysis, as conversely would introduce a bias into the dataset, where authors publishing in different subfields are classified according to different disambiguation procedures.

S6 Null-models for co-activities and transitions between subfields

In Fig. **??**d we map the relation between physics subfields into a network, where nodes represent subfields, and weighted links describe significant co-activity between them. Let us consider a set of N physicists, and two subfields α and β with respectively N^{α} and N^{β} physicists. We define the co-activity $C^{\alpha\beta}$ between the two subfields as the ratio between the number of physicists $N^{\alpha\beta}$ working on both subfields α and β, and the expected number $\hat{N}^{\alpha\beta} = (N^{\alpha}N^{\beta})/N$. Starting from the link with the highest weight, we plot the minimum number of links needed to have a connected network. All reported links have $C > 1$, meaning that only edges with co-activity higher than what expected at random (given the size of the subfields) are shown.

In Fig. **??**b we show flows of physicists from the subfield(s) of their first publication, to the subfield(s) where their activity is significant (RCA>1). Let us consider the number of physicists $F^{\alpha|\beta}$ working in subfield α who started their career by publishing in subfield β, so that $\sum_{\beta} F^{\alpha|\beta} = N^{\alpha}$. Subfield β is significantly contributing to subfield α only if $F^{\alpha|\beta}/N^{\alpha}$ is greater than the total fraction of physicists whose first publication is in subfield β (reported in the rectangles on the top). Only significant flows are shown.

S7 Subfields boundaries and overlaps

In the PACS classification more subfields can be assigned to a single paper. To what extent the measured authors' coactivity (Fig. **??**d) is simply due to these multiple assignments? To answer this question, we selected the papers published by authors that are active on (at least) two subfields, and measured how many papers reported at the same time PACS from both activity subfields. We found that only 19.4% of all papers have PACS from multiple subfields, with the remaining ∼81% being labelled with only one subfield. Fig. S9 shows the fraction of double-subfield papers for all pairs of subfields, indicating that on average the measured author coactivity is not due to multiple assignments at the level of single papers, but to authors working on truly different topics of research. Yet, despite the overall accuracy of our method, admittedly, there are examples of topics in which the classification we used is inadequate to fully allocate them in a single subfield. This is the case, for example, of network science, where 94% of the papers are consistently tagged with more than one subfield (mainly General, CondMat and/or Interdisc).

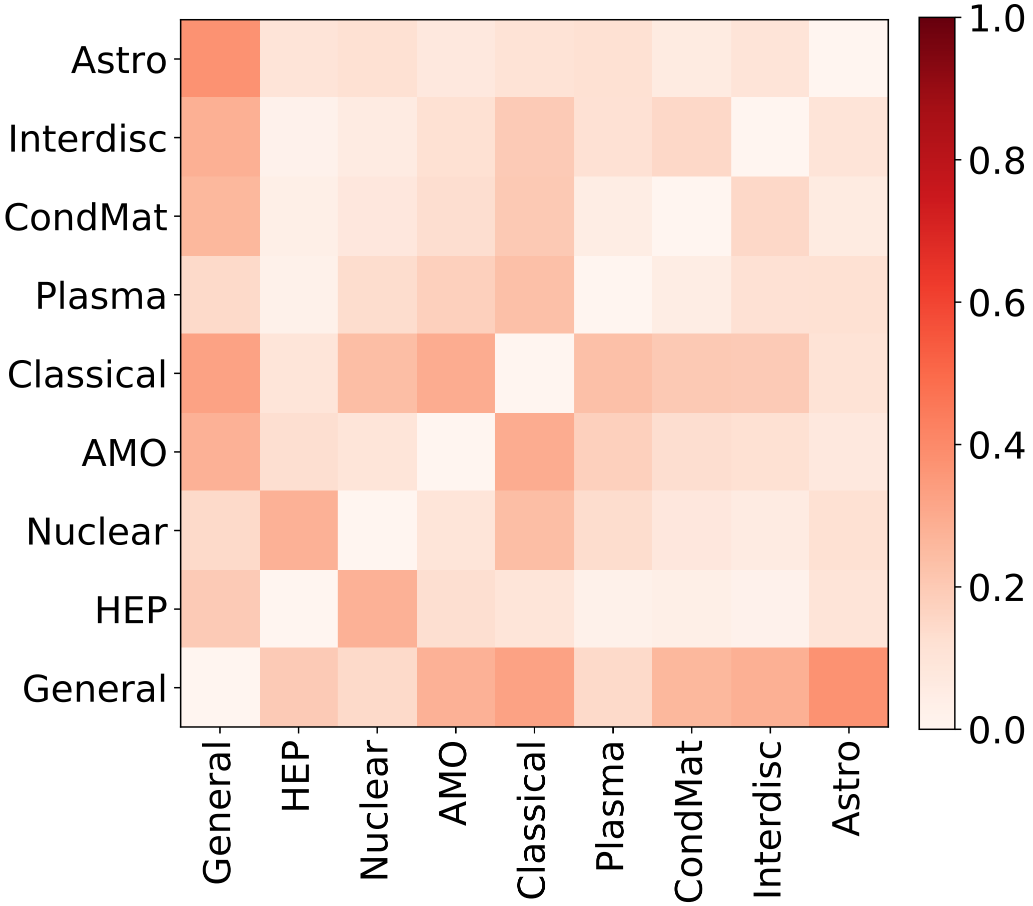

Figure S9. Subfields overlap in papers for non specialised physicists. The element i, j of the matrix represents the fraction of papers that are assigned to both subfields i and j.

S8 Theorists and experimentalists in HEP

In Fig. **??**a-e we show the different patterns of productivity and impact within different subfields. However, in the case of HEP it is worth separating the communities of experimentalists and theorists, to the different expected patterns of activity across the two groups. To do so we classified as "theoretical" the authors predominantly publishing in HEP with less than 10 co-authors. In contrast, the remaining ones were classified as "experimental", as a large average number of co-authors is likely to be due to involvement in large-scale collaborations, such as ATLAS or CMS at LHC. We validated our classification by looking at the occurrence of the strings “experiment*” and “theor*” in the titles of the two groups of papers. We found that the string “theor*” is $\sim$20 times more likely to be assigned to the theoretical group rather than to the experimental one (when compared to a null model that takes into account the different sizes of the two groups). On the other side, the string “experiment*” is $\sim$4 times more likely to appear in the experimental group than in the theoretical one. Fig. S10 shows the equivalent of Fig. **??**a-e for these two categories within HEP. Although experimentalists write significantly more papers and get more citations than theorists, the picture is clearly reversed when one considers fractional productivity and impact. Besides, it is clear that the overall peculiar behaviour in productivity and impact of HEP physicists is mainly due to the experimental part of the community. HEP theorists' productivity and impact are mostly comparable to physicists active in the other physics subfields.

S9 LHC and the HEP 2010 peak

In Fig. **??**a we show over the years the relative number of new authors entering each subfield. We notice that HEP is characterised by a large peak in 2010. For this reason we looked at all the first publications of new HEP authors in 2010, and searched for the collaborations responsible for each paper. We found that 76% of the new HEP authors in 2010 have a first publication which is connected to the opening of LHC, either directly through the ATLAS, CMS and LHC collaborations.,[13] or indirectly (Ref.[14] of the ALICE collaboration takes advantage of results by LHC). These new authors also amount to the 21% of the total number of new physicists across subfields, explaining the observed peak for HEP. In Fig. S11 we show the yearly fraction of physicists who published their first paper in a new subfield, after removing all new 2010 HEP authors connected to the activities of LHC. As displayed, the peak at 2010 for HEP disappears.

S10 Chaperone effect

In Fig. **??** we computed the number of chaperoned authors across subfields. The Chaperone effect was originally investigated in Ref.[15] for scientific venues, measured in terms of scientists making the transition from non-last to last (senior / PI) authors in papers published in a journal. Here, as we are interested in the relations, as well as migration between physics subfields, we focused on a simplified version of such chaperone measure c, computing the fraction of

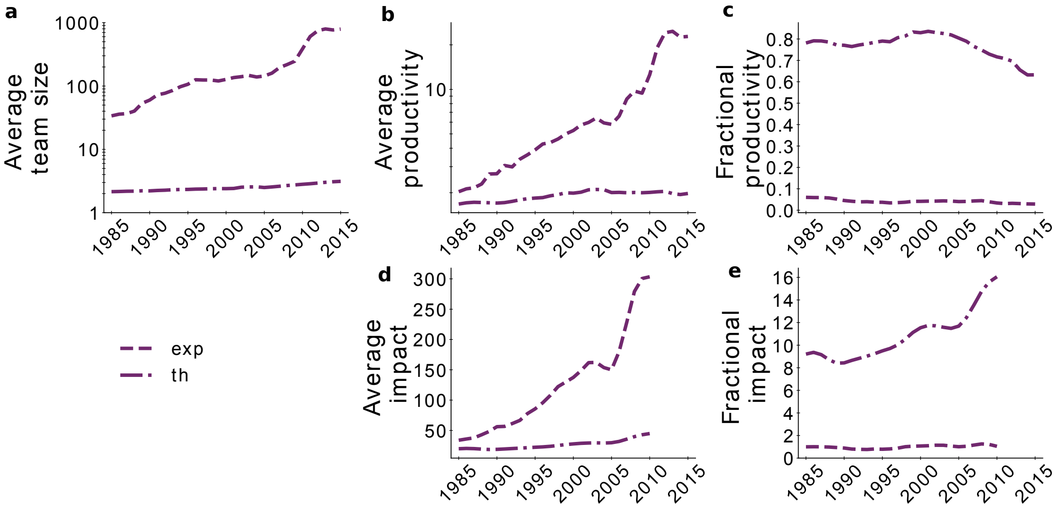

Figure S10. Productivity and impact for theorists and experimentalists in HEP**. a**, Average team size, defined as average number of authors per paper, over time. **b**, Average productivity, defined as average number of papers per author, over time. Productivity grows dramatically for experimentalists but stays roughly constant for the other group. **c**, Fractional productivity, i.e. number of papers divided by team size, over time. Here the picture is reversed: fractional productivity is significantly higher for theorists, although slightly declining over time. **d**, Average impact, defined as number of citations per author within a 5 years window. Impact increases in both groups, but for the experimentalists the growth is exceptional. **e**, Fractional impact, i.e. number of paper citations divided by team size, over time. Again, theorists get significantly more citations per author than experimentalists, with the difference also growing over time.

physicists first publishing in a subfield who have as co-authors at least one scientist who has already published in the area.

Despite being intuitive and close to the variable used in Ref.,[15] this measure might not prove adequate in the case of subfields characterised by publication through large-scale collaborations. For such a reason, we tested our results against $\tilde{c}$, a variant of the chaperone index. Given the first publication of a scientists in a subfield, $\tilde{c}$ measures the average fraction of co-authors who have already published in the area. As shown in Fig. S12, in the case of our data c and $\tilde{c}$ are very highly correlated, with a cosine similarity of 0.99.

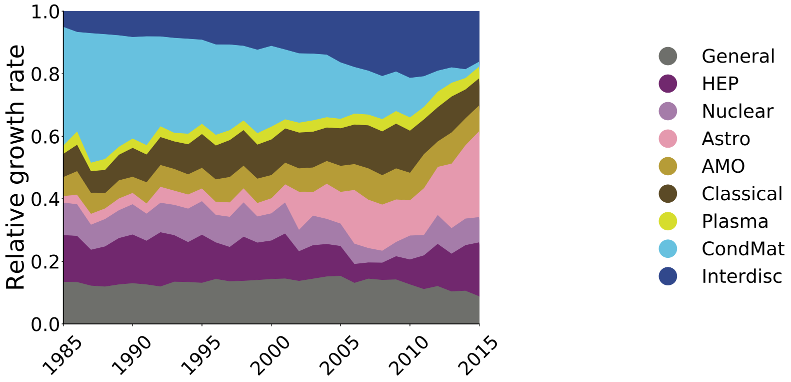

Figure S11. Relative growth rate of subfields after removing new 2010 HEP authors connected to the activities of LHC. No peak is observed for HEP authors in 2010. Colours for subfields are consistent with those used in the main text.

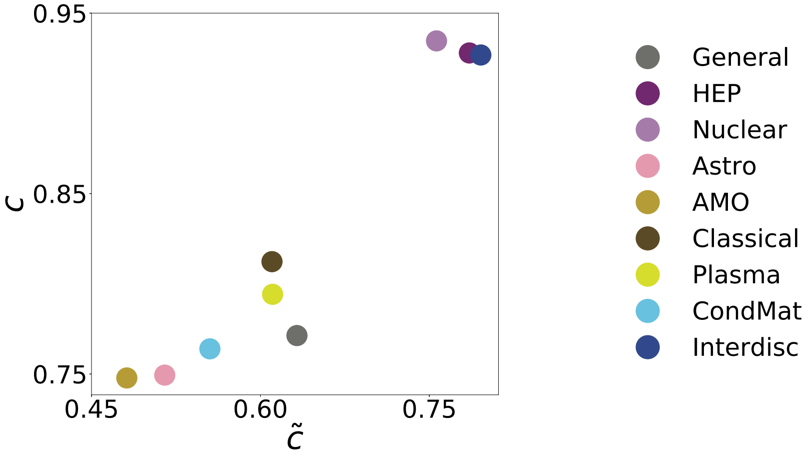

Figure S12. Comparison between two measures of Chaperone effect. Scatterplot between the original measure c to quantify the number of chaperoned authors, and the fractional measure $\tilde{c}$. The values of two variables across subfields in our dataset are highly correlated. Colours for subfields are consistent with those used in the main text.

S11 Authors impact and citation rates across subfields

Top authors across subfields have very different impact, as shown in Fig. **??**g. This is mainly a consequence of different productivities, rather than diverse citation patterns across subfields. Indeed, the typical number of papers produced by top authors is very heterogenous across physics communities (Fig. **??**f). In contrast, we found that the number of citations per paper is rather constant across subfields: the average is 27.3, with all subfields falling within 1.8 standard deviation from this value. For example, papers published in HEP and Interdisc receive on average respectively 27.4 and 33.9 citations, despite the much larger impact of HEP authors. Similar results are obtained for the medians of paper citations across subfields. The average median across physics communities is 9.0, the standard deviation of the median across subfields is 1.1, and all subfields are at most 1.7 standard deviation away from the global median. The median of paper citations for HEP and Interdisc are respectively 9 and 11.

S12 The physics Nobel prizes

In Fig. **??**j we show the distribution of Nobel prizes awarded in physics across subfields. Data on Nobel prizes in physics are available on the Nobel prize website.[16] We report all awards since

1985 in order to be consistent with the rest of our data-driven analysis of careers in physics. All such awards are accompanied by a motivation which allows to assign the crucial discovery or stream of research that led to the Nobel prize to one or more physics subfields. In the considered time span (1985-2017), 82 scientists were awarded the Nobel prize in physics.

References

1. Sinatra, R., Deville, P., Szell, M., Wang, D. & Barabási, A.-L. A century of physics. *Nature Physics* **11**, 791 (2015).

2. Deville, P. *Understanding social dynamics through big data (PhD Thesis)* (Université Catholique de Louvain, 2015).

3. Zhu, X. & Ghahramani, Z. Learning from labeled and unlabeled data with label propagation (2002).

4. PACS 2010 regular edition. https://publishing.aip.org/publishing/pacs/pacs-2010-regular-edition.

5. APS Dataset. https://journals.aps.org/datasets.

6. Palchykov, V., Gemmetto, V., Boyarsky, A. & Garlaschelli, D. Ground truth? concept-based communities versus the external classification of physics manuscripts. *EPJ Data Science* **5** (2016).

7. Balassa, B. Trade liberalization and 'revealed' comparative advantage. *Manchester School* 99–123 (1965).

8. Hidalgo, C. A. & Hausmann, R. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences* **106**, 10570–10575 (2009). URL http://www.pnas.org/content/106/26/10570. http://www.pnas.org/content/106/26/10570.full.pdf.

9. Aps divisions. https://www.aps.org/membership/units.

10. Smalheiser, N. R. & Torvik, V. I. Author name disambiguation. *Annual Review of Information Science and Technology* **43**, 1–43 (2009). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aris.2009.1440430113. https://onlinelibrary.wiley.com/doi/pdf/10.1002/aris.2009.1440430113.

11. Most common chinese surnames. https://en.wikipedia.org/wiki/List_of_common_Chinese_surnames.

12. Most common korean surnames. https://en.wikipedia.org/wiki/List_of_Korean_surnames.

13. Yetkin, T. New physics at atlas and cms experiments with the first data. *Nuclear Physics B - Proceedings Supplements* **200-202**, 17 – 26 (2010). URL http://www.sciencedirect.com/science/article/pii/S0920563210000836. The International Workshop on Beyond the Standard Model Physics and LHC Signatures (BSM-LHC).

14. Aamodt, K. *et al.* Midrapidity antiproton-to-proton ratio in pp collisons at $\sqrt{s} = 0.9$ and 7 tev measured by the alice experiment. *Phys. Rev. Lett.* **105**, 072002 (2010). URL https://link.aps.org/doi/10.1103/PhysRevLett.105.072002.

15. Sekara, V. *et al.* The chaperone effect in science. *PNAS, in print* (2018).

16. Physics Nobel prizes. https://www.nobelprize.org/prizes/uncategorized/all-nobel-prizes-in-physics.