Entropic selection of concepts in networks of similarity between documents

Andrea Martini,1,2 Alessio Cardillo,1,2 and Paolo De Los Rios1

1Laboratory for Statistical Biophysics, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
2Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, E-50018 Zaragoza, Spain

Scientists have devoted many efforts to study the organization and evolution of science by leveraging the textual information contained in the title/abstract of scientific documents. However, only few studies focus on the analysis of the whole body of a document. Using the whole text of documents allows, instead, to unveil the organization of scientific knowledge using a network of similarity between articles based on their characterizing concepts which can be extracted, for instance, through the ScienceWISE platform. However, such network has a remarkably high link density (36%) hindering the association of groups of documents to a given topic, because not all the concepts are equally informative and useful to discriminate between articles. The presence of “generic concepts” generates a large amount of spurious connections in the system. To identify/remove these concepts, we introduce a method to gauge their relevance according to an information-theoretic approach. The significance of a concept $c$ is encoded by the distance between its maximum entropy, $S_{\text{max}}$, and the observed one, $S_c$. After removing concepts within a certain distance from the maximum, we rebuild the similarity network and analyze its topic structure. The consequences of pruning concepts are twofold: the number of links decreases, as well as the noise present in the strength of similarities between articles. Hence, the filtered network displays a more refined community structure, where each community contains articles related to a specific topic. Finally, the method can be applied to other kind of documents and works also in a coarse-grained mode, allowing the study of a corpus at different scales.

The science of science, namely the scientific study of scholar activities, leverages the availability of large amounts of data to provide a new and quantitative view of the dynamical organization of the scientific community and of its activities. It has been possible, for example, to study the patterns of citations between research articles [1]–[3], the structure of scientific collaborations [4]–[5], their stratification and geographical distribution [6]–[9], and to identify the best contributions and most successful actors [10]–[13].

Unfortunately, the wealth of data that makes the science of science possible is turning, in recent times, into a serious issue for scientists [14]–[16]. It has become clear that, in order to stay up to date with the advances within a given discipline, reading all the newly published documents would require an excessive amount of time, enforcing the decision to read only those documents that can be considered of relevance.

To assist researchers in such selection process, several tools have been developed throughout the years [17]. Most of them make use of the meta-information attached to the documents (title, abstract, keywords, references and so on) to recommend selected contents. However, the amount of information available in a title, or an abstract, may not be enough to identify the topic a given document belongs to. A better approach is to perform a semantic analysis on a document and extract its relevant concepts [18].

The classification of documents according to their main “topic” has attracted the interest of the scientific community [19]–[22]. One of the most recent attempts to create such classification is the ScienceWISE platform (SW); whose aim is: “to create a suite of convenient web-based tools that will encourage scientists to organize the scientific information into an ontology-like structures and make scientific papers more self-contained”. To map the ontological structure of a collection of documents, they can then be considered as the nodes of a network, and the weight of their connections captures the similarity between their characterizing concepts.

However, the presence of “common” concepts appearing in almost every document results in a network which is very dense, akin to an almost complete graph [23]–[24]. The SW platform tries to make the pruning of “common concepts” easier by letting expert users tag them. Nevertheless, the manual curation of common concepts/buzzwords requires the allocation of a considerable portion of time for the users – assuming their willingness to cooperate. Also, the massive amount of documents, often from domains that are only weakly related (as, e.g., subdomains in physics), demands the presence of a large number of experts with vastly different competences. Furthermore, what can be considered common for an expert in a context may not be so for others, leading to ambiguities. Hence, the definition of common concepts tout-court without any objective approach may lead to biases and errors. Given these premises, an automatic filtering method able to discriminate common concepts based on objective, measurable observables would be highly desirable.

In the present manuscript, we propose an approach toward the solution of these problems. More specifically, we design a method that – given a set of documents – identifies generic concepts according to their entropies. This, in turn, allows the pruning of spurious connections between documents due to generic concepts, fostering the emergence of the underlying topic structure of the network. After introducing the method, we apply it on a collection of physics articles as well as on a collection of web texts on climate change. By performing community detection on the filtered network, we identify specific topics in a way that goes beyond a broad area classification, like arXiv categories. Our findings highlight the fact
that being common is an attribute of a concept that strongly depends on the context of the collection under study, and that is non-trivially associated to its frequency within the collection itself.

I. SIMILARITY NETWORKS AND CONCEPT FILTERING

The recent advent of “big data” has triggered a flurry of activity in many disciplines such as computational social science and science of science [25, 26]. The availability of detailed metadata information attached to publication records constitutes an authentic treasure trove. Information like date, title, abstract, affiliations, keywords and bibliographies can, among other things, be used to classify manuscripts into topics by means of semantic analysis techniques [19, 20, 22].

In this paper, we consider the network of similarity between manuscripts extracted using the SW platform (see Materials and Methods for details). We consider two distinct datasets: scientific manuscripts submitted to the arXiv e-print archive in the Physics section and web texts on climate change. Given a corpus of documents \( N_a \), each document is parsed and its concepts are automatically extracted and weighted according to their relevance. The set of concepts pertaining to document \( \alpha \) is denoted by \( C_\alpha \). Each document is then considered as a node of a network, and a link connects two nodes if there is some similarity between their contents, i.e. if they share at least one concept. The weight of the link between two nodes \( \alpha \) and \( \beta \), \( w_{\alpha\beta} \in [0,1] \), estimates the extent of such similarity.

One naive attempt to gauge the similarity of a document \( \alpha \) with respect to the whole corpus is by counting its number of connections, i.e. its degree, \( k_\alpha \). By computing the average degree \( \langle k \rangle \) in both datasets (Tabs. I and Svii) we immediately see that their values are of the order of the size of the corpus \( N_a \) itself. This means that, on average, one article is similar to all the others present in the corpus. Such density is due to the presence of the so-called “common concepts” (hereafter CC), which spuriously increase the weights of the connections between documents that are not, otherwise, really similar. Therefore, the widespread presence of CC is responsible for the loss of one of the major advantages of framing the system as a network, i.e. sparsity.

Several solutions to the network sparsification problem have been proposed [27–29]. All of them ensure the conservation of the statistical properties of the original network acting in an ex-post way. A more suitable approach, instead, would be to act ex-ante, directly on the process responsible for the generation of the weights. This translates into acting directly on the concepts to select only the relevant ones before computing the weights.

The SW platform has a built-in list of CC that has been prepared with the collaboration of users who are expert in Physics. In particular, these users can either tag as common those concepts that are already present on the platform, or suggest/recommend new ones. Obviously, updating the CC list requires the active cooperation of users. This task could become more taxing, given the amount of documents and concepts to validate, and the rate at which they are deposited. More importantly, the tag of CC relies solely on the verdict of experts and does not take into account the topic composition of the corpus under scrutiny. As an example, the concept graphene could be considered as common within a corpus composed mainly of articles about Material Science. Instead, it should possibly be treated as a specific one in a corpus focused on Astrophysics. Thus, simply removing concepts that are manually declared as common might be inappropriate. The aforementioned naive example highlights the weaknesses of the current approach and, thus, calls for an alternative solution to CC tagging.

Can we design a method to automatically select relevant concepts which also accounts for the composition of the collection? To find an answer to this question, we must first sketch the identikit of our “ideal/relevant concept”. A relevant concept must fulfill at least two prerequisites: i) It must neither be too frequent (i.e. be a buzzword) or too rare (too specific) in the corpus. ii) It must appear a significant number of times within a document. The first property accounts for the ability of a concept to discriminate between documents while being statistically relevant. The second property regards the concept relevance in the characterization of the content of a document. Given a corpus with \( N_c \) documents, and being \( C = \bigcup_{\alpha=1}^{N_a} C_\alpha \) the set of all its concepts, the discriminative power of a concept \( c \in C \) appearing in \( N_c \) documents is its document frequency \( df_c = \frac{N_c}{N_a} \). Its relevance for a given document \( \alpha \) is measured as the term frequency \( tf_c(\alpha) \), which is the number of times \( c \) appears in \( \alpha \). The average term frequency of \( c \) is \( \langle tf_c \rangle = \frac{1}{N_a} \sum_{\alpha=1}^{N_a} tf_c(\alpha) \).

Concepts are thus characterized in terms of their \( df \) and \( \langle tf \rangle \). We tessellate the corresponding two-dimensional space by imposing thresholds on \( df \) and \( \langle tf \rangle \) to delimit regions where ubiquitous, relevant and rare concepts fall. The resulting tessellation is displayed in Figs. [S1] and [S6]. Although being very intuitive and straightforward, imposing thresholds on \( df \) and \( \langle tf \rangle \) is problematic for several reasons. Concepts that have been manually tagged as common in the Physics dataset do not fall in any particular location in the \( df/ \langle tf \rangle \) plane (black diamonds in Fig. [S6]). At best, they tend to follow a law that is not a simple combination of \( df \) and/or \( \langle tf \rangle \). Furthermore, as seen in [30–34] the probability that a word appears inside a text or a corpus \( n \) times tends to follow the Zipf’s law [51] or, in general, to be broad, as we show in Fig. [S5]. Hence, imposing a characteristic scale on scale-free quantities is not only subjective, but likely right away incorrect. These limitations call for alternative ways to filter concepts based on their microscopic behaviour.

Based on Shannon notion of information [36], we can associate the importance of a concept to its entropy, \( S_c \), [37, 39], defined as

\[
S_c = - \sum_{t=1}^{\max(tf_c)} q_c(t) \ln q_c(t),
\]

\(^2\)http://arxiv.org
where \( q_c(t) = \frac{N_c(t)}{N(t)} \) is the probability of finding a document where concept \( c \) has \( tf = t \). It is worth mentioning that if the length of the documents in a corpus is not constant the same approach could still be used considering the \( tf \) density, \( \frac{rtf}{t} = \frac{tf}{L} \), where \( L \) is the length of the document. Interestingly, concepts hand-marked as common tend to have \( S_c \) higher than others with the same value of \( \langle tf \rangle \) (Fig. 1 panel A), and tend to visually accumulate toward an ideal null of the distribution of the concepts in the \( (S_c, \langle tf \rangle) \) plane. The same behavior, even more pronounced, is observed in the case of \( \langle ln(tf) \rangle \) (see SI, Fig. S7). This observation suggests that there could be some underlying mechanism (highlighted by entropy) that pushes common concepts toward their maximum possible value. We have thus decided to check if a similar behavior could be reproduced with a Maximum Entropy Principle (MEP) approach [34, 39] and we have computed for each concept its maximum entropy \( S_{max}(c) \) constrained by the values of \( \langle tf_c \rangle \) and \( \langle ln (tf_c) \rangle \) (see Materials and Methods). Indeed, direct inspection for several concepts, and in particular for the ones manually tagged as common, reveals that \( q_c(t) \) is well described by a power-law with cutoff, \( q_c(t) \propto t^{-s}e^{-\lambda t} \) (see insets in Fig. 2), which is precisely the functional form expected from a maximum entropy principle with the above mentioned constraints.

Remarkably, the MEP approach reveals that \( s = 3/2 \) is the most frequent value of the power-law exponent (Fig. 2), which is the exponent typical of critical branching processes [40]. Although further investigations in this direction go beyond the scope of the present work, it is suggestive to picture the appearance of papers in the arXiv as a process where older manuscripts "inspire" ("generate", in the branching process language) new papers containing a similar number of concepts, each appearing a similar number of times.

Using entropies as a new set of coordinates, we arrange concepts on the \( (S_c, S_{max}) \) plane as reported in Figs. 1 and ST1. As expected, the majority of common concepts lay close to the \( S_c = S_{max} \) line. We exploit then this feature to design a new criterion to discriminate concepts. For each concept, we define its residual entropy \( S_R(c) \) as the difference between its maximum and measured entropies \( S_R(c) = S_{max}(c) - S_c(c) \). Such quantity is – up to a scaling factor – equivalent to the Kullback-Leibler divergence between the observed probability distribution, \( q_c(t) \), and the maximum entropy one, \( p_c(t) \), reported in Eq. 5 [41] (see STD of SI for details). We then assign concepts to different percentiles \( p \) of the probability distribution of \( S_R \). The color of the dots in Fig. 1 panel C, accounts for the value of \( p \) and the lower inset is the projection of the percentile information on the \( (df, \langle tf \rangle) \) space. Finally, we can use \( p \) as a sort of "distance" from the maximum entropy curve \( S_c = S_{max} \), and thus consider as significant those concepts having \( p \geq \tilde{p} \), and use them to build the similarity network.

A closer inspection of Fig. 1 panel C, reveals two interesting facts. On the one hand, there are outlier concepts that were marked as common by SW experts but that are located far from the \( S_c = S_{max} \) line. Examples are 'operational calculus', 'Fraunhofer line', 'gigawatt' and 'Gaussian symplectic ensemble' which could be thought as generic only within a very selected range of topics, but become quite specific in a corpus spanning a wider range of topics as in our dataset (see Fig. S4). On the other hand, there are many concepts close to the diagonal but not marked as common such as 'statistics', 'intensity', 'Hamiltonian', 'fluid dynamics', and 'scaling law' that have slipped through the attention of experts without being tagged as generic albeit being obviously so. Both cases highlight the weak points of the human based concept tagging as well as the advantages of our entropy based criterion which is, to a certain extent, almost unsupervised.

It is worth remarking that entropy has already been used in natural language processing to quantify the relevance of words within single texts [34, 42, 44]. However, apart from focusing on single documents, these studies seek to understand the role played by the position of words within texts. More importantly, none of them use the difference of entropies to assess the relevance of words in discriminating the content of documents within a collection which, in turns, constitutes the cornerstone of our approach. Our ranking method differs also from another well know approach of natural language processing, namely the ranking based on the Inverse Document Frequency (IDF) [45, 46], albeit the two quantities are not completely unrelated (see Sec. SIIA.3 of SI). Finally, as shown in Tabs. STI–STII, our method can also be used to gauge the relevance of a concept within a given document.

II. RESULTS

This entropy-based objective criterion allow us discarding concepts before using them to construct the similarity network between documents, thus realizing the sought-for \textit{ex-ante} approach. The consequences of concept filtering on the topology of similarity networks are displayed in Tab. I and STII. In the case of the Physics dataset, the total number of concepts \( N_{con} = |C| \), as well as the number of documents having at least one concept \( N_c \) decrease with \( p \), albeit the latter remains pretty constant up to \( p = 30\% \). The link density \( \rho \), instead, is dramatically affected. It goes from 36\% to 7\% when \( p \) goes from 0\% to 10\%. Consequently, the maximum and average degrees \( k_{max} \) and \( \langle k \rangle \) drop significantly with \( p \), while the average distance between documents \( \langle d \rangle \) increases. The increase of \( \langle d \rangle \) with \( p \) – together with the fragmentisation into distinct connected components \( M \) – is the byproduct of the existence of “cultural holes” among distinct topics of Physics and, more in general, science itself [47]. A similar trend can be observed also for the climate dataset (see Tab. STIV). The sparsification of the connection pattern implies that spurious connections are less abundant. As a direct consequence, documents on the same topic will look more akin and, hence, tend to form more tightly connected groups. Finally, the mesoscale organization of the network in groups/topics could be obtained using a \textit{community detection} method [21, 48].
The lower inset is the projection of the percentile information on the plane. The color of the points encodes the percentile regression between the maximum entropy computed fixing only $S$. The solid line is the linear least-squares regression between $\log(df)$ and $S_c$ for CC (Pearson correlation coefficient $r = 0.743$). (C) Organization of the concepts in the $(S_c, S_{\max})$ plane. The lower inset is the projection of the percentile information on the $(\log(df), \log(tf))$ plane.

FIG. 1. Relations between entropy and other features for the concepts of the Physics dataset. SW common concepts (CC) are represented as black diamonds. (A) Relation between the entropy, $S_c$, and the average value of the term-frequency, $(tf)$. The dashed line corresponds to the maximum entropy computed fixing only $(tf)$. (B) Relation between the $df$ and the entropy $S_c$. The solid line is the linear least-squares regression between $\log(df)$ and $S_c$ for CC (Pearson correlation coefficient $r = 0.743$). (C) Organization of the concepts in the $(S_c, S_{\max})$ plane. The color of the points encodes the percentile $p$ of the residual entropy distribution $P(S_d)$ to which concepts belong (concepts with $p > 90\%$ are omitted). The solid line is the $S_c = S_{\max}$ curve. The Pearson correlation coefficient between $S_{\max}$ and $S_c$ for CC is $r = 0.979$.

**TABLE I.** Topological indicators of the Physics preprint’s similarity networks. The row $p = 0\%$ corresponds to the original network, while $p > 0\%$ to networks filtered using the maximum entropy. In the columns we report: the percentage of filtered concepts $p$, the number of concepts $N_{\text{con}}$, of documents having at least one concept $N_d$, the link density $\rho$, the average $(k)$ and maximum degrees $k_{\text{max}}$, the average clustering coefficient $(C)$, the average path length $(L)$ and the number of connected components $M$.

| $p$ (%) | $N_{\text{con}}$ | $N_d$ | $\rho$ (%) | $(k)$ | $k_{\text{max}}$ | $(C)$ | $(L)$ | $M$ |
|---------|---------------|-------|------------|-------|----------------|-------|-------|----|
| 0       | 11637         | 52979 | 36.493     | 19333.522 | 46504 | 0.557 | 1.635 | 1   |
| 10      | 9594          | 52337 | 7.340      | 3841.235  | 17532 | 0.327 | 1.935 | 1   |
| 20      | 8528          | 51522 | 3.752      | 1933.031  | 10399 | 0.319 | 2.008 | 1   |
| 30      | 7462          | 49821 | 2.057      | 1024.818  | 8109  | 0.332 | 2.160 | 1   |
| 40      | 6396          | 47173 | 1.197      | 564.823   | 5669  | 0.343 | 2.378 | 2   |
| 50      | 5330          | 41775 | 0.638      | 266.419   | 2771  | 0.390 | 2.687 | 7   |
| 60      | 4505          | 39084 | 0.363      | 89.766    | 1140  | 0.755 | 3.409 | 59  |
| 70      | 3197          | 24710 | 0.228      | 13.027    | 104   | 0.848 | 7.124 | 342 |
| 80      | 2047          | 14035 | 0.133      | 4.124     | 70    | 1.124 | 7.124 | 342 |
| 90      | 1066          | 5703  | 0.228      | 13.027    | 104   | 0.848 | 7.124 | 342 |

A. Organization of documents into topics

Knowing the topic structure of a corpus is of utmost importance in platforms like Amazon, aNobii or Reddit where recommendations rely on the successful classification of documents according to their contents. In the case of ScienceWISE and Physics documents, inferring the topic structure has at least two possible implications. On the one hand, it could be used to recommend contents to users or to cross-validate the PhySH system recently adopted by the American Physical Society to classify manuscripts. On the other hand, it can be used to portray the coarse grained process of specialization undergone in Physics and in all Science in general. To this aim, we study the evolution of the community structure of the corpus when we progressively reduce the pool of concepts using our entropic filtering. The result of the analysis is reported in the Sankey/alluvial diagram of Figs. 3 and S12.

In such diagram, each box represents a community and its height is proportional to the number of documents in the community. Each column refers to a different filter intensity $p$. The name of each community refers to the main topic (see Materials and Methods). The evolution of the community structure reveals some intriguing and interesting features. When $p = 0$, the detected communities clearly correspond to the major topics in physics. No finer grouping is possible at this

---

1. [https://www.amazon.com/](https://www.amazon.com/)
2. [https://www.reddit.com/](https://www.reddit.com/)
3. [http://www.anobii.com/](http://www.anobii.com/)
4. The interactive version of these diagrams displaying additional information is available at [50](http://www.anobii.com/)
level because of the presence of common concepts, that act as a "glue" within large communities. As \( p \) increases, there is a progressive fragmentation of topics passing from broad areas of Physics – not exactly overlapping with the \textit{arXiv} classification as shown by [S2] – to increasingly specific subjects at \( p = 20\% \). An example is the fragmentation of Astrophysics (\( p = 0\% \)) into Stellar Physics, Planetology, High Energy Astrophysics and Solar Physics, which progressively develops up to \( p = 40\% \).

Although the pruning of common concepts allows to detect smaller and more specific communities, pushing it to overly large values of \( p \) deteriorates the results. This is due to a combination of different effects. On the one hand, when too many common, or even mildly common concepts are eliminated, connections between papers end up being determined by concepts whose statistical significance is poor. On the other hand, a growing fraction of papers simply disconnect from the network ("ghost" papers). Heuristically, filtering should not proceed beyond a level \( p_{\text{opt}} \) such that the size of the ghost-papers group exceeds the average size of all the other communities. For the Physics dataset \( 30\% \leq p_{\text{opt}} \leq 40\% \) while for climate we have \( 25\% \leq p_{\text{opt}} \leq 30\% \) as shown in Fig. S15.

We have also repeated the same analysis that we have described here for the Physics corpus from the \textit{arXiv} on a set of web documents about climate change. In this case, we had to slightly modify the approach by considering the probability distribution of the rescaled \( tf \), namely the \( tf \) of concept \( c \) in paper \( \alpha \) (\( tf_c(\alpha) \)) divided by the length of the paper (\( L(\alpha) \)), i.e. \( rtf_c(\alpha) = \frac{tf_c(\alpha)}{L(\alpha)} \), because of the presence in the set of groups of documents of vastly different size. Correspondingly, the entropy had to be redefined as an integral (see [SIC 2] of SI for methodological details, Secs. SII B 2 and SII B 3 for the results). Furthermore, the climate dataset was parsed for keywords; without an ontology structure. As a consequence the similarity between documents is less precise. This notwithstanding, filtering based on the entropy of keywords can still be used to generate a document network that is more tractable for community detection techniques. At the same time, the identification of concepts of different degrees of generality might be used to generate an ontology for this document corpus.

In general, the phenomenology of filtering can be grouped into three classes: i) preservation with specialization (e.g. Cond_Mat/Astrophysics) \textit{i.e.} when the topic of a community remains unaltered but the concepts used to characterize it are more specific; ii) splitting with specialization (e.g. Cond_Mat \( \rightarrow \) Graphene + Solid State) \textit{i.e.} when the removal of generic concepts ends up in the fragmentation of the original topic into more specific sub-topics; iii) nucleation (e.g. Extreme weather in climate dataset) \textit{i.e.} when one or more shared concepts draw documents together into a single community. Finally, we want to stress that the community structure obtained by ranking concepts with the \( IDF \) and residual entropy \( S_d \) are different especially for values of \( p \) higher than 10\%, as shown in Figs. S9 and S10.

\[ rtf_c(\alpha) = \frac{tf_c(\alpha)}{L(\alpha)} \]

\section{III. Discussion}

The access to the semantic content of whole documents grants an unprecedented opportunity for their classification and can improve the search of contents within huge collections. However, such opportunity comes at a hefty price: the similarity networks are extremely dense, hindering the retrieval of their topic structure/landscape. Buzzwords or common concepts are responsible for such excessive density. In the present manuscript we have presented a method based on maximum entropy to automatically select relevant concepts to filter the similarity networks and improve the topic modeling of big document corpora. According to the method, common concepts are those whose entropy is closer to their maximum one. As a consequence, the definition of common stemming from our method is more “liquid” than the one used by the SW platform since it does not rely on user validation and, more importantly, depends on the content of the documents under scrutiny. By selectively pruning a fraction \( p \) of concepts, we obtain an improved topic description of two different corpora: scientific preprints on Physics and web documents on climate change (Fig. 3 and S12). Finally, the entropic filtering proposed here could be applied in a recursive way on sub-corpora of documents or can be used to study the evolution in time of the generality of a concept (like graphene or Python). Last but not least, the method could be used also to improve already existing ontologies.
IV. MATERIALS AND METHODS

A. Data

We consider two collections of documents. One containing scientific manuscripts from arXiv\(^5\), a repository of electronic preprints of scientific articles, and another made of web articles on climate change extracted using the underlying machinery of the ScienceWISE platform. In the case of scientific manuscripts, we selected documents submitted during year 2013 under the physics categories either as primary or secondary subjects resulting in a corpus of 52,979 articles (Tab. I). The composition of the corpus in terms of arXiv categories is reported in Tab. [S1]. The climate change corpus has been built selecting web documents written in English with at least 500 words, whose URLs are mentioned by – at least – 20 distinct tweets (see Sec. [S1B.1] of SI). Text are parsed and keywords are extracted using KPEX algorithm \(^5\). After that, keywords are matched with concepts available in a crowdsourced ontology accessible on the platform. The ontology has been built by initially collecting scientific concepts from online encyclopedias and subsequently refined with manual inspection by experts. The second step is missing for climate web documents since no ontology is available in SW. The Physics dataset is composed by 11,637 concepts, from which we discarded those appearing always with the same \(tf\) ending up with 10,661 concepts, 339 of which have been marked as “common” by SW. For the climate dataset, instead, we have 152,871 keywords. By deleting those having a \(\text{max}(rtf) − \text{min}(rtf) ≤ 0.005\), only 9222 keywords are left.

B. Similarity between documents

The resemblance between pairs of documents is measured using the cosine similarity of their concept vectors \(\vec{d}\). Given the set of concepts used in a corpus with \(N_d\) documents, \(C = \bigcup_{\alpha=1}^{N_d} C_\alpha\), the relevance of a concept \(c\) in a document is given by its \textit{term frequency} \(tf_c\). For each document \(\alpha\), we consider a concepts vector \(d_\alpha\), whose elements correspond to the TF-IDF of its concepts, a standard measure used in information retrieval \(\text{45, 46} \). Hence, we have:

\[
d_\alpha(c) = \begin{cases} 
    tf_c(\alpha) \cdot \text{IDF}_c & \text{if } c \in C_\alpha, \\
    0 & \text{otherwise},
\end{cases}
\]  

\(^5\text{https://arxiv.org/}\)
where $t_f c (\alpha)$ is the boosted term frequency, i.e. the number of times $c$ appears in $\alpha$ modulated according to the location (title, abstract, body) where it appears. The other factor, $IDF_c$, is the Inverse Document Frequency and accounts for the frequency with which a concept appears in the corpus. In particular, $IDF$ penalizes concepts used frequently since:

$$
IDF_c = \ln \left( \frac{N_a}{N_c} \right),
$$

where $N_c$ is the number of papers that contain concept $c$. The weight $w_{\alpha \beta}$ of the link encoding the similarity between two documents $\alpha$ and $\beta$ is equal to:

$$
w_{\alpha \beta} = \frac{\vec{d}_\alpha \cdot \vec{d}_\beta}{\| \vec{d}_\alpha \| \| \vec{d}_\beta \|},
$$

where $\cdot$ denotes the scalar product and $\| \cdot \|$ is the Euclidean norm. The weight of a link falls in the interval $[0, 1]$, where $w = 0$ indicates documents sharing any concept at all (i.e. are completely different), and whose vectors form an angle $\theta = 90^\circ$. A value $w = 1$ is found if the documents not simply have the same set of concepts but use them in the same way (i.e. are identical) thus having vectors forming an angle $\theta = 0^\circ$. If $w_{\alpha \beta}  \leq w_{\text{min}}$ we do not add the link between $\alpha$ and $\beta$ to the similarity network. We set $w_{\text{min}} = 0.01$ corresponding to an angle $\theta_{\text{min}} = 89.43^\circ$, ensuring that we are pruning only connections encoding spurious similarities. Finally, we remark that using the $tf$ or its density, $rtf$, to compute weights is equivalent.

1. Maximum entropy principle

To gauge how informative a concept can be, we calculate (using Eq. [1]) its entropy $S_c$ based on the term-frequencies $t_f c$. We have observed that concepts labeled as common in the SW platform tend to have a higher entropy with respect to other concepts having the same $(tf)$. To corroborate such regularity, we have applied the maximum entropy principle to the distribution of the term-frequencies of a concept, $t_f c$, to determine the associated probability mass function that satisfies certain constraints. As shown in Supplementary Information, Sec. [STC] the selection of the empirical values of the first moment and log-moment, $(t_f c)$ and $(\ln (t_f c))$ [33], as constraints implies a probability mass function of the following form:

$$
p_c(t) = \frac{e^{-\lambda t}}{\text{Li}_s(e^{-\lambda})},
$$

where $\text{Li}_s(e^{-\lambda})$ is the polylogarithm of order $s$ and argument $e^{-\lambda}$, defined as:

$$
\text{Li}_s(e^{-\lambda}) = \sum_{t=1}^{\infty} \frac{e^{-\lambda t}}{t^s}.
$$

The parameters $\lambda$ and $s$ are determined, for each concept $c$, imposing the constraints to Eq. [5] and solving numerically the system of equations:

$$
\text{Li}_{s-1}(e^{-\lambda}) = (t_f c),
$$

$$
- \partial_s \text{Li}_s(e^{-\lambda}) = \langle \ln (t_f c) \rangle.
$$

As a consequence, the maximum entropy $S_{\text{max}}$ is:

$$
S_{\text{max}} = \ln \left[ \text{Li}_s(e^{-\lambda}) \right] + \lambda \langle t_f c \rangle + s \langle \ln (t_f c) \rangle.
$$

2. Community detection

To detect the community structure of the similarity network we used the Louvain method [54], a well known greedy algorithm that finds the optimal partitioning of the system by maximizing a function named modularity introduced by Newman in [55].

ACKNOWLEDGMENTS

All the authors acknowledge the financial support of SNSF through the project CRSII2_147609. The authors also thank Alex Constantin for many helpful discussions and the developer team of ScienceWISE for its help.

[1] de Solla Price DJ (1965) Networks of scientific papers. Science 149(3683):510–515.
[2] Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: Toward an objective measure of scientific impact. Proc Natl Acad Sci USA 105(45):17268–17272.
[3] Kuhn T, Pere M, Helbing D (2014) Inheritance Patterns in Citation Networks Reveal Scientific Memes. Phys. Rev. X, 4:041036.
[4] Newman MEJ (2004) Coauthorship networks and patterns of scientific collaboration. Proc Natl Acad Sci USA 101(Suppl 1):5200–5205.
[5] Milojević S (2014). Principles of scientific research team formation and evolution. Proc Natl Acad Sci USA, 111:3984–9.
[6] Jones BF, Wuchty S, Uzzi B (2008) Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. Science 322, 1259–1262.
[7] Grauwin S, Jensen P (2011) Mapping scientific institutions. Scientometrics 89, 943–954.
[8] Gargiulo F, Carletti T (2014) Driving forces of researchers mobility. Scientific Reports 4:4860.
[9] Kumar Pan R, Kaski K, Fortunato S (2012) World citation and collaboration networks: uncovering the role of geography in science. Sci Rep 2:902
[10] Chen P, Xie H, Maslov S, Redner S (2007) Finding scientific gems with Google’s PageRank algorithm. Journal of Informetrics 1:8–15.

[11] Ke Q, Ferrara E, Radicchi F, Flammini A (2015) Defining and identifying Sleeping Beauties in science. Proc Natl Acad Sci USA 112(24):7426–7431.

[12] Wang D, Song C, Barabási A-L (2013) Quantifying long-term scientific impact. Science 342:127–32.

[13] Petersen AM, et al. (2014) Reputation and impact in academic careers. Proc Natl Acad Sci USA 111(43):15316–15321.

[14] Van Noorden R (2014) Global scientific output doubles every nine years. Nature News Blog. Available at: http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html

[15] Bornmann L, Mutz R (2015) Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. J Assn Inf Sci Tec 66(11): 2215–2222.

[16] Ginsparg P (2011) ArXiv at 20. Nature 476(7359):145–7.

[17] Gibney E (2014) How to tame the flood of literature. Nature 513 (7516):129–30.

[18] Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. J. Mach. Learn. Res. 3:993.

[19] Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci USA 101(Suppl 1):5228–35.

[20] Mane KK, Börner K (2004) Mapping topics and topic bursts in PNAS. Proc Natl Acad Sci USA 101 Suppl(1):5287–90.

[21] Lencichinetti A, et al. (2015) High-reproducibility and high-accuracy method for automated topic classification. Phys Rev X 5(1):011007.

[22] Silva FN, et al. (2016) Using network science and text analytics to produce surveys in a scientific topic. Jour Informetrics 10(2):487–502.

[23] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang, D U (2006) Complex networks: Structure and dynamics. Phys Rep 424, 175–308.

[24] Newman MEJ (2010) Networks. Oxford University Press

[25] Lynch C (2008) Big data: How do your data grow? Nature 455:28–29.

[26] Lazer D et al. (2009). Computational Social Science. Science 323:721–723.

[27] Serrano MA, Boguñá M, Vespignani A (2009) Extracting the multiscale backbone of complex weighted networks. Proc Natl Acad Sci USA 106(16):6483–6488.

[28] Radicchi F, Ramasco JJ, Fortunato S (2011) Information filtering in complex weighted networks. Phys Rev E 83:046101.

[29] Gemmetto V, Cardillo A, Garlaschelli D (2017) Enhanced extraction of weighted networks backbones.

[30] Ferrer i Cancho R, Sole RV (2003) Least effort and the origins of scaling in human language. Proc Natl Acad Sci USA 100(3):788-791.

[31] Font-Clos F, Boleda G, Corral A (2013) A scaling law beyond Zipf’s law and its relation to Heaps’ law. New J. Phys. 15:093033.

[32] Gerlach M, Altmann E G (2014) Scaling laws and fluctuations in the statistics of word frequencies. New J. Phys., 16:113010.

[33] Visser (2013) Zipf’s law, power laws and maximum entropy. New J. Phys. 15:043021.

[34] Yan X-Y, Minnhagen P (2015) Maximum Entropy, Word-Frequency, Chinese Characters, and Multiple Meanings. PLoS ONE 10(5): e0125592.

[35] Zipf GK (1949) Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology. (Cambridge, MA: Addison–Wesley).

[36] Shannon, C E (1948) A Mathematical Theory of Communication. Bell System Technical Journal 27(3): 379–423.

[37] Berger A, Della Pietra S, Della Pietra V A (1996) A Maximum Entropy Approach to Natural Language Processing. Comput. Linguist. 22:39–71.

[38] Hotho A, Nünberger A, Paas G (2005) A Brief Survey of Text Mining. LDV Forum – GLDV Journal for Computational Linguistics and Language Technology 20:19–62.

[39] Back S K, Bernhardsson S, Minnhagen P (2011) Zipf’s law unzipped. New J. Phys. 13:043004.

[40] Harris TE (1963) The Theory of Branching Process. (Berlin, Springer).

[41] Kullback S, Leibler R A (1951) On information and sufficiency. Ann. Math. Stat. 22(1):79–86

[42] Carpena P, Bernaola-Galván PA, Carretero-Campos C and Coronado AV (2016) Probability distribution of intersymbol distances in random symbolic sequences: Applications to improving detection of keywords in texts and of amino acid clustering in proteins. Phys Rev E 94(5):052302

[43] Herrera J P, Purdy P A (2008) Statistical keyword detection in literary corpora Journal. Eur. Phys. J. B 63(1):135–146

[44] Altman EG, Dias L, Gerlach M (2017) Generalized Entropies and the Similarity of Texts. J. Stat. Mech. 014002

[45] Sparck Jones K (1972) A statistical interpretation of term specificity and its application in retrieval. Journal of documentation 28(1):11–21.

[46] Robertson S (2004) Understanding inverse document frequency: on theoretical arguments for IDF. Journal of documentation 60(5):503–520.

[47] Vilhena D, et al. (2014). Finding Cultural Holes: How Structure and Culture Diverge in Networks of Scholarly Communication. Sociological Science, 1:221–238.

[48] Fortunato S, Hric D (2016). Community detection in networks: A user guide. Phys Rep, 659:1–44.

[49] Physics Subject Headings (PhySH). Available at: https://physsh.aps.org/

[50] Interactive Sankey diagrams, Available at: http://www.bifi.es/~cardillo/data.html#semantic

[51] Rosvall M, Bergstrom C T (2010) Mapping change in large networks. PLoS ONE, 5:e8694.

[52] Palchovsky V, Gemmetto V, Boyarsky A, Garlaschelli D (2016) Ground truth? Concept-based communities versus the external classification of physics manuscripts. Eur. Phys. J Data Science, 5:28.

[53] Constantin A (2014) Automatic Structure and Keyphrase Analysis of Scientific Publications. PhD Thesis, The University of Manchester.

[54] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech 2008:P10008.

[55] Newman M E J (2006) Modularity and community structure in networks. Proc. Natl Acad. Sci. USA 103:8577–82.
Supplementary Materials for the manuscript entitled: Entropic selection of concepts in networks of similarity between documents

SI. THEORY

In this section we provide the theoretical details behind our maximum-entropy based filtering method. We begin introducing the two-dimensional tessellation filtering (Sec. SI A). Then, we prove the relation between full entropy, \( S_f \), and conditional one, \( S_c \), and we motivate why we based the filtering methodology on the conditional entropy instead of the full one (Sec. SI B). After that, we provide the details of the maximum entropy models used in the main text (Sec. SI C) and we demonstrate the equivalence between the residual entropy, \( S_d \), and the Kullback-Leibler divergence between the probability distributions of empirical observations and maximum entropy model (Sec. SI D). Finally, we present the comparison between the concept lists ranked according to residual entropy \( S_d \) and \( IDF \) (Sec. SI E 1) and between communities after filtering these concept lists (Sec. SI E 2).

A. Two-dimensional tessellation

As we have commented in the main text, the document frequency \( df \) and average term frequency \( \langle tf \rangle \) (or its average density \( \langle rtf \rangle \)), are two features that could be used to characterize concepts. More specifically, we can use such features as coordinates of a two-dimensional space, thus classifying a concept \( c \) by the position it occupies in the \( \langle tf_c \rangle - df_c \) plane. We tessellate the plane by imposing thresholds on the coordinates to delimit regions where ubiquitous, rare and relevant concepts fall as shown in Fig. S1.

![Fig. S1. Tessellation of the \( \langle tf_c \rangle - df_c \) plane used to classify concepts. Each colored area defines one type of concept. Logarithmic scale is used to visualize neatly the results. The dashed lines are used to delimit the tiles.](image)

We define the following tiles on the \( \langle tf_c \rangle - df_c \) plane:

**A1:** The domain of specific/rare concepts characterized by having small values of both \( df \) and \( \langle tf \rangle \).

**A2:** The domain of common/ubiquitous concepts displaying high values of both \( df \) and \( \langle tf \rangle \).

**A3:** The domain of relevant/informative concepts corresponding to those having intermediate values of \( df \) and \( \langle tf \rangle \).

**A4:** All the remaining concepts appearing within documents not enough times (on average) to be considered relevant.

If we consider each trait separately, we can divide its space into three – or more – domains denoting low, medium and high values of such trait. Specifically, we consider two values for \( \langle tf_c \rangle \) (25% and 75%), and three for \( df_c \) (25%, 75%, and 90%), instead. We use percentiles values since both quantities tend to have a broad distribution, hence raw values are not suitable to highlight their variability. Therefore, we can estimate the similarity between documents using only those concepts classified as relevant. Nevertheless, despite its intuitive nature, the tessellation method is not a good filtering approach since it depends on too many parameters, whose numbers and values are arbitrary. Moreover, as shown later on in Fig. S6, the tessellation is unable to reproduce all the features of those concepts tagged as generic by ScienceWISE. Such limitations drove us to abandon the two-dimensional filtering in favor of alternative approaches.
B. Relation between full entropy and conditional entropy

The probabilistic formulation of entropy used in the main text does not contemplate as an event the absence of a concept $c$ in a document (i.e., $tf = rf = 0$). For this reason, in Eq. [1] the sum (integral) starts from $tf = 1$ $(rtf = \varepsilon > 0)$. Such entropy, $S_c$, gets labeled as conditional since it is computed under the condition that concept $c$ appears in the document. However, it is possible to define another probability distribution including the absence event which translates into another entropy, $S_f$, labeled as full. To construct such distribution, we consider the total number of papers in the collection, $N_a$, while concept $c$ appears only in $N_c \leq N_a$. Then, we extend the $tf_c$ probability distribution by incorporating the absence event as a term that corresponds to the fraction of papers where the concept $c$ did not appear. Such term is exactly $1 - df_c$, where $df_c = \frac{N_c}{N_a}$ is nothing else that the document frequency of concept $c$. In conclusion, the probability for the concept $c$ appearing $k \in [0, \infty]$ times is then $q_c(k) = \frac{N_c(k)}{N_a}$, where $q_c(0) = 1 - df_c$. Therefore, the full entropy associated to distribution $q_c(k)$ is:

\[
S_f = - \sum_{k=0}^{\infty} q_c(k) \ln q_c(k) = - q_c(0) \ln q_c(0) - \sum_{k=1}^{\infty} q_c(k) \ln q_c(k) = -(1 - df_c) \ln(1 - df_c) - \sum_{k=1}^{\infty} \frac{N_c(k)}{N_a} \ln \left( \frac{N_c(k)}{N_a} \right).
\]

Since $\frac{N_c}{N_a} = \frac{N_c(k)}{N_c} \cdot \frac{N_c}{N_a}$, we have:

\[
S_f = -(1 - df_c) \ln(1 - df_c) - \sum_{k=1}^{\infty} \frac{N_c(k)}{N_c} df_c \ln \left( \frac{N_c(k)}{N_c} \right) = -(1 - df_c) \ln(1 - df_c) - df_c \sum_{k=1}^{\infty} \frac{N_c(k)}{N_c} \ln \left( \frac{N_c(k)}{N_c} \right) = S_{bin} + df_c \cdot S_c.
\]

where we used the normalization condition over the $N_c(k)$, i.e., $\sum_{k=1}^{\infty} \frac{N_c(k)}{N_c} = 1$. The full entropy $S_f$ is a linear combination of two entropies: the binary entropy, $S_{bin}$, and the conditional entropy, $S_c$, respectively. The first accounts for the probability of presence/absence of a concept in the collection. The second is the entropy computed in Eq. [1] of the main text but modulated by the $df_c$.

It is natural to ask whether $S_f$ could be used to classify concepts as well as $S_c$, or not. To this aim, in Fig. S2 we display the relation among $S_f$ and several quantities to assess if $S_f$ is a valid alternative to $S_c$ in discriminating generic concepts. More specifically, in panel (A) we display the relation between $(tf)$ and $S_f$ -- in analogy with Fig. [1] (A). The comparison of the two figures strikingly highlights the ability of $S_f$ to grasp the tendency of generic concepts to display higher entropies for a given value of $(tf)$. Following the parallelism with Fig. [1] in panel (B) we inspect the relation between $S_f$ and $df$ confirming the inability of the former to discern generic concepts, alike to what displayed in Fig. [1] B. One may argue that the quantities used so far -- $(tf)$ and $df$ -- are the most naive ones, and there could be others better suited to improve the performances of $S_f$. However, the adoption of quantities as: $S_c$ (panel C), $df \cdot S_c$ (panel D) and $\frac{df \cdot S_c}{S_f}$ (panel E) does not change situation: a clear separation between generic concepts and the others is missing. Finally, we perform the same analysis for the first term in Eq. S1, $S_{bin}$. The full entropy $S_f$ does not present a characteristic dependence for the generic concepts either on $S_{bin}$, panel (F), or on its fraction explained by $S_{bin}$, $\frac{S_{bin}}{S_f}$, panel (G). In a nutshell, none of the relations displayed in Fig. S2 seems to provide additional clues to design a classification criterion. Hence, the full entropy $S_f$ is unfit to distinguish generic concepts, since its discriminative power is weaker than the conditional one.

C. Maximum entropy models

The maximum entropy principle provides a framework to compare the amount of information carried by concepts, as encoded by the conditional entropy $S_c$. However, the raw value of $S_c$, which quantifies the actual information present in the data, is not enough to establish if a concept is generic or not. Indeed, we need to fairly assess if the observed entropy $S_c$ is small or not when confronted to an expected value. Such theoretical counterpart of the observed entropy is the maximum entropy and it is associated to a theoretical distribution where some features are fixed. The required features are quantities extracted from the data and the maximum entropy distribution is then the maximally random distribution that reproduce such features. In such a way, the established features uniquely determine the maximum entropy distribution. Operationally, in order to impose the constraints
FIG. S2. Relation between the full entropy $S_f$ and other quantities. We consider: average term frequency $\langle tf \rangle$ (A), document frequency $df$ (B), conditional entropy $S_c$ (C), the contribution of conditional entropy to full one $df \cdot S_c$ (D), conditional entropy contribution over the full one $\frac{df}{S_f} \cdot S_c$ (E), binary entropy $S_{bin}$ (F) and binary entropy contribution over the full one $\frac{S_{bin}}{S_f}$ (G).

derived from the fixed features, we adopt the Lagrangian multipliers formalism, a tool that allow to easily establish the maximum entropy distribution that fulfills the constraints. In the rest of the section, we describe two different maximum entropy models with the respective features and we detail the calculations that lead to the associated distributions.

1. Discrete TF

The first maximum entropy model is devised to characterize the observed term-frequency distribution of a concept, $tf_c$. The probability that concept $c$ appears $k$ times in the collection is simply the fraction of papers where it is present $k$ times, i.e. $q_c(k) = \frac{N_c(k)}{N}$. Since the most typical feature of a distribution is its mean value, it is reasonable to consider the average term-frequency $\langle tf_c \rangle$ as a constraint of the maximum entropy distribution. Furthermore, in the literature there have been many evidences that the term-frequency distribution spans several orders of magnitude and has a fat tail profile (see references in the main text). However, to properly describe the observed behavior, we have to include another constraint which is the average logarithm of the term-frequency, $\langle \ln(tf_c) \rangle$. We denote the expected probability of concept $c$ occurring $k$ times as $p_c(k)$. The analytical form of $p_c$ is then calculated by maximizing its entropy $S_{max}$ under the constraints on the average term-frequency and the average logarithm of the term-frequency, which must be equal to $\langle tf_c \rangle$ and $\langle \ln(tf_c) \rangle$ respectively:

$$S = -\sum_{k=1}^{\infty} p_c(k) \ln p_c(k) + \lambda \left( \langle tf_c \rangle - \sum_{k=1}^{\infty} k p_c(k) \right) + s \left( \langle \ln(tf_c) \rangle - \sum_{k=1}^{\infty} \ln(k) p_c(k) \right) + \nu \left( 1 - \sum_{k=1}^{\infty} p_c(k) \right). \quad (S2)$$

In the above equation, $\lambda$ is the Lagrange multiplier associated to the constraint $\langle tf_c \rangle$, $s$ is the one associated to $\langle \ln(tf_c) \rangle$ and $\nu$ is associated to the normalization condition of the probability mass function $p_c(k)$. The maximization of Eq. S2 with respect to $p_c(k)$ is performed as $\frac{\partial S}{\partial p_c(k)} = 0$, which gives:

$$- \ln p_c(k) - 1 - \lambda k - s \ln(k) - \nu = 0. \quad (S3)$$

Thus, the probability mass function $p_c$ is defined as:

$$p_c(k) = \frac{e^{-(\nu+1)} e^{-\lambda k}}{k^s}. \quad (S4)$$
This probability mass function corresponds to a power law with a cutoff. The power law \( k^{-s} \) is responsible for the fat tail of the distribution, while the cutoff \( e^{-\lambda k} \) is likely due to the finite size of the collection of articles under scrutiny. The maximization of Eq. \([S2]\) with respect to each Lagrangian multiplier allows to impose the respective constraint. In turn, such constraints determine the parameters that appear in Eq. \([S4]\). Thus, maximizing Eq. \([S2]\) with respect to \( \nu \), \( \frac{\partial S}{\partial \nu} = 0 \), we recover the normalization condition:

\[
\sum_{k=1}^{\infty} \rho_c(k) = e^{-(\nu+1)} \sum_{k=1}^{\infty} \frac{e^{-\lambda k}}{k^s} = 1,
\]

\[
e^{(\nu+1)} = \sum_{k=1}^{\infty} \frac{e^{-\lambda k}}{k^s} = \text{Li}_s(e^{-\lambda}).
\]

(S5)

In the last equation, the summation is equal to the special function called polylogarithm of order \( s \) and argument \( e^{-\lambda} \). For any value of \( s, e^{-\lambda} \in \mathbb{C} \), the validity of such expression is limited to the case when the modulus of the argument is smaller than one, \(|e^{-\lambda}| < 1\). However, in the present case we are interested only on real valued parameters. Eq. \([S5]\) allows to properly normalize the probability mass function in Eq. \([S4]\), so that we obtain:

\[
\rho_c(k; s, \lambda) = \frac{e^{-\lambda k}}{k^s \text{Li}_s(e^{-\lambda})}.
\]

(S6)

The maximization of Eq. \([S2]\) with respect to \( \lambda \), \( \frac{\partial S}{\partial \lambda} = 0 \), allows to express the constraint \( \langle tf_c \rangle \):

\[
\sum_{k=1}^{\infty} k \rho_c(k) = \sum_{k=1}^{\infty} \frac{k e^{-\lambda k}}{\text{Li}_s(e^{-\lambda})} = \frac{\sum_{k=1}^{\infty} e^{-\lambda k}}{\text{Li}_s(e^{-\lambda})} = \langle tf_c \rangle,
\]

\[
\frac{\text{Li}_{s-1}(e^{-\lambda})}{\text{Li}_s(e^{-\lambda})} = \langle tf_c \rangle.
\]

(S7)

Note that in the last equation we used the definition of the polylogarithm and the normalization constant obtained in Eq. \([S5]\).

Finally, the constraint on \( \langle \ln(tf_c) \rangle \) is imposed by maximizing Eq. \([S2]\) with respect to \( s \), \( \frac{\partial S}{\partial s} = 0 \):

\[
\sum_{k=1}^{\infty} \ln(k) \rho_c(k) = \sum_{k=1}^{\infty} \ln(k) \frac{e^{-\lambda k}}{k^s} = \langle \ln(tf_c) \rangle,
\]

\[
- \frac{\partial_s \text{Li}_s(e^{-\lambda})}{\text{Li}_s(e^{-\lambda})} = \langle \ln(tf_c) \rangle.
\]

(S8)

To derive the expression in the last line we used the identity:

\[
\sum_{k=1}^{\infty} \frac{\ln(k) e^{-\lambda k}}{k^s} = - \frac{\partial_s}{\partial s} \sum_{k=1}^{\infty} \frac{e^{-\lambda k}}{k^s} = - \frac{\partial_s}{\partial s} \text{Li}_s(e^{-\lambda}).
\]

From Eqs. \([S7]\) and \([S8]\) we see that both parameters \( \lambda \) and \( s \) are present in each of them. Since they are coupled in both equations they cannot be retrieved in an explicit form but we have to solve Eqs. \([S7]\) and \([S8]\) simultaneously through a numerical method.

The details of the algorithmic implementation of the system, along with some snippets of code, are provided in \([SIII B 1]\). The maximum entropy \( S_{max} \) associated to the probability in Eq. \([S6]\) is then:

\[
S_{max} = - \sum_{k=1}^{\infty} \rho_c(k) \ln \rho_c(k)
\]

\[
= \ln \left[ \text{Li}_s(e^{-\lambda}) \right] + \lambda \langle tf_c \rangle + s \langle \ln(tf_c) \rangle.
\]

(S9)

2. Density of TF

The second maximum entropy model is conceived to represent the rescaled version of the term-frequency distribution of a concept \( c \), denoted as \( rtf_c \). In particular, the density of the term-frequency accounts for the length \( L(\alpha) \), in terms of words, of
document $\alpha$ where concept $c$ is present, $rtf_c(\alpha) = \frac{tf_c(\alpha)}{L(\alpha)}$. The term-frequency density is better suited to describe the relevance of a concept within documents when their length is inhomogeneous. In the opposite case documents exhibit a typical length scale and the usage of raw term-frequency $tf_c$ is more appropriate since it does not alter the observed frequency.

Being the term-frequency density a continuous variable we have to adopt a probability density function $p_c(x)$ to define its maximum entropy distribution. The two constraints that we set are the average and variance of the logarithm of the term-frequency density, $\langle \ln(rtf_c) \rangle$ and $\sigma^2(\ln(rtf_c))$. We take the logarithm of the term-frequency density since it is more appropriated to describe a broad distribution of values: the average of the logarithm identifies the most likely value of the distribution while the variance characterize its variability scale. The analytical expression of the probability density function $p_c(x)$ is determined by maximizing its entropy $S_{\text{max}}$ under the constraints on the average and variance of the logarithm of the term-frequency density that must equate $\langle \ln(rtf_c) \rangle$ and $\sigma^2(\ln(rtf_c))$ respectively:

$$S = - \int_0^\infty p_c(x) \ln p_c(x) \, dx$$

$$+ \gamma \left( \langle \ln(rtf_c) \rangle - \int_0^\infty \ln(x) p_c(x) \, dx \right)$$

$$+ \eta \left[ \sigma^2(\ln(rtf_c)) - \int_0^\infty \left( \ln(x) - \int_0^\infty \ln(x) p_c(x) \, dx \right)^2 p_c(x) \, dx \right]$$

$$+ \nu \left( 1 - \int_0^\infty p_c(x) \, dx \right).$$

(S10)

In Eq. (S10) we introduced the Lagrange multipliers $\gamma$, $\eta$ and $\nu$ that are correspondingly associated to the constraints $\langle \ln(rtf_c) \rangle$, $\sigma^2(\ln(rtf_c))$ and the normalization condition of the probability density function $p_c(x)$. From the maximization of Eq. (S10) with respect to $p_c(x)$, $\frac{\partial S}{\partial p_c(x)} = 0$, we obtain:

$$- \ln p_c(x) - 1 - \gamma \ln(x) - \eta (\ln(x) - \mu)^2 - \nu = 0,$$

(S11)

where we defined the constant $\mu = \int_0^\infty \ln(x) p_c(x) \, dx$, which is the average value of the logarithm of $x$ according to the maximum entropy distribution $p_c(x)$. As a consequence, the probability density function $p_c$ is defined as:

$$p_c(x) = \frac{e^{-(\nu+1)\ln(x)-\eta(\ln(x)-\mu)^2}}{x^\gamma}.$$

(S12)

As we did in the previous case, Sec. (S11) we must impose the normalization condition on the probability density function Eq. (S12), i.e. the analogous of Eq. (S5), and we must calculate the parameters $\eta$ and $\gamma$, similarly to what we performed in Eqs. (S7) and (S8). Since we have already detailed the process to calculate the parameters in Sec. (S11) we do not report here the intermediate steps but directly the full expression of the probability density function:

$$p_c(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi \sigma} x} \exp \left[ - \frac{(\ln x - \mu)^2}{2 \sigma^2} \right] \quad \text{with } x > 0.$$

(S13)

Such probability density function corresponds to the lognormal that describes the distribution of a positive random variable $x$ whose logarithm $\ln(x)$ follows a normal distribution. Thanks to the constraints imposed on the average and the variance of the logarithm of $x$, the parameters $\mu$ and $\sigma^2$ that appear in Eq. (S13) can be directly calculated from the observed data:

$$\mu = \int_0^\infty \ln(x) p_c(x) \, dx \equiv \langle \ln(rtf_c) \rangle, \quad \sigma^2 = \int_0^\infty (\ln(x) - \mu)^2 p_c(x) \, dx \equiv \sigma^2(\ln(rtf_c)).$$

(S14)

Note that $\mu$ is a constant determined from the actual data and is not a function of $\sigma^2$. In contrast, in Eqs. (S7) and (S8) the parameters $\lambda$ and $s$ were coupled. The maximum entropy $S_{\text{max}}$ associated to the probability in Eq. (S13) is then:

$$S_{\text{max}} = - \int_0^\infty p_c(x) \ln p_c(x) \, dx$$

$$= \ln \left( \sqrt{2\pi \sigma} \right) + \mu + \frac{1}{2}.$$

(S15)
D. Equivalence between Kullback-Leibler divergence and entropy difference, $S_d$

In main text, for each concept $c$, we defined its residual entropy, $S_d(c)$, as the difference between the maximum entropy $S_{max}(c)$ and the conditional entropy $S_c(c)$ (unless explicitly indicated, we avoid to specify concept $c$ in the notation). Here we show that $S_d$, used to characterize the generality of the concepts, is exactly equivalent to the Kullback-Leibler divergence (KL), a widely used measure used to compute the difference between two probability distributions \cite{Cover2012}. More specifically, we consider the KL between the maximum entropy probability distribution $p$ (see Eq. 5) and the empirical observed one $q$. For the sake of simplicity, we demonstrate here the case where $p$ and $q$ are probability mass functions describing discrete random variables, although the same reasoning can be used in the case of probability density functions. In particular, we recall that the observed probability that the $t$-f of a given concept is equal to $k$, which corresponds to $N(k)$, where $N(k)$ is the number of papers where the concept appears $k$ times and $N_a$ is the total number of papers. The Kullback-Leibler divergence from $p$ to $q$ is defined as:

$$D_{KL}(q||p) = \sum_{k=1}^{\infty} q(k) \ln \left( \frac{q(k)}{p(k)} \right) = - \sum_{k=1}^{\infty} q(k) \ln p(k) + \sum_{k=1}^{\infty} q(k) \ln q(k).$$  \hspace{1cm} (S16)

The last term in the Eq. [S16] is nothing else, apart for the sign, than the conditional entropy $S_c$:

$$S_c = - \sum_{k=1}^{\infty} q(k) \ln q(k).$$  \hspace{1cm} (S17)

The first term, instead, can be rewritten using the maximum entropy probability $p$ (see Eq. 5) as:

$$- \sum_{k=1}^{\infty} q(k) \ln p(k) = - \sum_{k=1}^{\infty} q(k) \ln \left( \frac{e^{-\lambda k}}{\text{Li}_s(e^{-\lambda})} \right) = \sum_{k=1}^{\infty} q(k) \ln \left[ \text{Li}_s(e^{-\lambda}) \right] - \sum_{k=1}^{\infty} q(k) \ln \left( \frac{e^{-\lambda k}}{k^s} \right)$$

$$= \ln \left[ \text{Li}_s(e^{-\lambda}) \right] - \sum_{k=1}^{\infty} q(k) \ln \left( \frac{e^{-\lambda k}}{k^s} \right) = \ln \left[ \text{Li}_s(e^{-\lambda}) \right] + \lambda \langle k \rangle + s \langle \ln k \rangle \equiv S_{max}.$$

Plugging the results of Eqs. [S17] and [S18] into Eq. [S16] we get:

$$D_{KL}(q||p) = - \sum_{k=1}^{\infty} q(k) \ln p(k) + \sum_{k=1}^{\infty} q(k) \ln q(k) = S_{max} - S_c.$$  \hspace{1cm} (S19)

Hence, for a given concept $c$, the KL divergence of between $p$ and $q$ coincides with the residual entropy $S_d = S_{max} - S_c$.

E. Comparisons between sets

1. Comparison between concepts’ rankings based upon residual entropy or IDF

Despite being both defined on the collection under scrutiny, the IDF and the residual entropy $S_d$ encode different information. One penalizes concepts present in many papers; the other is more intrinsically related to the distribution of the frequency of a concept. As a consequence, we expect to observe some correlation among them albeit being two different indicators. To compare the list of concepts ranked upon their residual entropy, $S_d$, or their inverse document frequency, $IDF$, we compute the overlap, $O_{n,m}$, between the set of concepts falling in the $n$-th slice of the percentile of $IDF$, $A_n$, and the set of concepts falling in the $m$-th slice of the percentile of $S_d$, $B_m$. By $k$-th slice of the percentile $\tilde{P}$ of a probability distribution $p(x)$, we refer to the range of values of $x$ such that $x \in [\tilde{P}_{k-10}, \tilde{P}_k]$, $k \in \{10, 20, \ldots, 90\}$. Thus, we have:

$$O_{n,m} = \frac{|A_n \cap B_m|}{|A_n|},$$  \hspace{1cm} (S20)

with $O_{n,m} \in [0, 1]$. As usual, $O_{n,m} = 1$ denotes complete overlap between the two sets, while $O_{n,m} = 0$ denotes that the sets have no element in common.
2. Comparison between communities

One of the standard measures used to compute the overlap between two communities of a networked system is the Jaccard score \( J \). Given a graph \( G \) with \( N \) nodes, several methods can be used to retrieve its organization into modules \[48\]. The Jaccard score between a pair of sets of nodes (i.e. modules) \( A \) and \( B \), \( J_{A,B} \), is given by:

\[
J_{A,B} = \frac{|A \cap B|}{|A \cup B|} \quad J_{A,B} \in [0, 1].
\] (S21)

A value of one denotes complete overlap between the two sets (i.e. the sets are the same), while a value of zero denotes that the sets completely different.

SII. DATASETS

In this section we provide the details of the collections of documents used in our study, namely: Physics 2013 (Sec. SII A) and climate change (Sec. SII B), respectively. Here, we anticipate that the maximum entropy model based on term-frequency \( tf_c \), described in Sec. SI C 1, will be used for the Physics 2013 collection, while the model based on term-frequency density \( r_{tf_c} \) will be adopted for the climate change collection. We decided to work with two different maximum entropy models since the distribution of the document length exhibits different traits for the two collections, as shown in Fig. S3. More specifically, the document length for Physics 2013 collection displays a typical value of \( \sim 3000 \) words, while for the climate change collection it does not present a characteristic scale and is quite inhomogeneous. In the following, for each collection of documents, we comment how the data have been collected and how the progressive pruning of concepts reveals their organization into topics.

![Fig. S3](image)

**FIG. S3.** Distribution \( P(L) \) of the number of words per document, \( L \), for Physics 2013 (A) and climate change (B) collections. The Physics case shows the clear presence of a typical length of the documents around \( L \sim 3000 \), while this is not the case for climate change. Panels (C) and (D) display the cumulative distribution functions \( P(\leq L) \) for the same collections.

A. Physics 2013

1. Source

The documents of the Physics dataset are scientific manuscripts submitted to the arXiv \[58\] archive, a repository of electronic preprints of scientific articles spanning several domains of science. We selected manuscripts submitted during year 2013 under the physics categories either as primary or secondary subjects resulting in a corpus of \( N_a = 52979 \) documents. In Tab. S1 and Fig. S4 we report the composition of the collection in terms of these main categories: physics (physics), condensed matter (cond-mat), astrophysics (astro-ph), quantum physics (quant-ph), mathematical physics (math-ph), nonlinear sciences (nlin),...
general relativity and quantum cosmology (gr-qc), nuclear physics (nucl) and high-energy physics (hep). The last two further divides into theory (nucl-th) and experiment (nucl-ex) for nuclear and theory (hep-th), phenomenology (hep-ph), lattice (hep-lat) and experiment (hep-ex) for high-energy. The interested reader can check [58] for a more detailed description of each category. From the donut chart (Fig. S4) it is clear that the collection is highly inhomogeneous with cond-mat and astro-ph categories summing together almost half of the entire collection, while gr-qc, nucl, math-ph and nlin not even reaching the 13%.

| Category | $N_a$ | %  |
|----------|-------|----|
| cond-mat| 12679 | 23.90 |
| astro-ph | 12458 | 23.48 |
| hep | 9661 | 18.22 |
| physics | 7407 | 13.96 |
| quant-ph | 4039 | 7.61 |
| gr-qc | 2273 | 4.28 |
| nucl | 1891 | 3.57 |
| math-ph | 1767 | 3.34 |
| nlin | 876 | 1.65 |
| Total | 52979 | 100 |

TABLE Si. Number of articles $N_a$ and percentual size, %, of each arXiv principal category in the Physics 2013 dataset. Categories are listed in descending order of size.

The topological characteristics of the similarity network obtained from this collection are reported in the first row ($p = 0\%$) of Tab. [Sii]. The other rows, instead, refer to the networks obtained removing a percentage $p$ of concepts using the entropic filtering.

| $p$ (%) | $N_{con}$ | $N_a$ | $\rho$ (%) | $\langle k \rangle$ | $k_{max}$ | $\langle C \rangle$ | $\langle l \rangle$ | $M$ |
|---------|-----------|-------|-------------|-------------------|----------|------------------|---------|-----|
| 0       | 11637     | 52979 | 36.493      | 19333522          | 46504    | 0.557            | 1.635   | 1   |
| 10      | 9594      | 52377 | 7.340       | 3841235           | 17532    | 0.327            | 1.935   | 1   |
| 20      | 8528      | 51522 | 3.752       | 1933031           | 10399    | 0.319            | 2.008   | 1   |
| 30      | 7462      | 49821 | 2.057       | 1024818           | 8109     | 0.332            | 2.160   | 1   |
| 40      | 6396      | 47173 | 1.197       | 564823            | 5669     | 0.343            | 2.378   | 2   |
| 50      | 5330      | 41775 | 0.638       | 266419            | 2771     | 0.390            | 2.687   | 7   |
| 60      | 4264      | 34939 | 0.482       | 168307            | 1999     | 0.508            | 2.914   | 20  |
| 70      | 3197      | 24710 | 0.363       | 89766             | 1140     | 0.755            | 3.409   | 59  |
| 80      | 2132      | 14789 | 0.257       | 37989             | 495      | 0.783            | 4.242   | 153 |
| 90      | 1066      | 5703  | 0.228       | 13027             | 104      | 0.848            | 7.124   | 342 |

TABLE Sii. Topological indicators of the Physics 2013 similarity networks. The first row ($p = 0\%$) corresponds to the original network, while the others to the networks obtained using the maximum entropy filter. In the columns we report: the percentage of filtered concepts $p$, the number of concepts $N_{con}$, the number of articles containing at least one concept (nodes) $N_a$, the link density $\rho$, the average and maximum degrees $\langle k \rangle$ and $k_{max}$, the average clustering coefficient $\langle C \rangle$, the average path length $\langle l \rangle$ and the number of connected components $M$. The minimum edge weight is equal to $w_{min} = 0.01$.

2. Two-dimensional tessellation

Although the entropic filtering outperforms the two-dimensional tessellation one, we have decided to compute anyway the position of concepts on the $(\langle tf \rangle, df)$ plane to check if such representation highlights some interesting features. First of all, we remark that the distributions of such quantities are both heterogeneous, as shown in Fig. [S5]. Thus, establishing a threshold on quantities that miss a characteristic scale is not properly justifiable. Then, in Fig. [S6] we report the position of concepts on the $(\langle tf \rangle, df)$ plane. More specifically, in panel (A) we colored concepts according to the class they belong to, as outlined in Sec. [SIA] 46% of them (the green circles) are thus defined as significant by the tessellation scheme. As a comparison, in panel (B) we color each concept according to the value of its residual entropy percentile $p$: apart from the stratification of the residual entropy percentiles, we are unable to spot any other significant features.

The conditional entropy $S_c$ used in the entropic filtering is able to capture a peculiar trend of the common concepts (CC) not only when examined against the average term-frequency $\langle tf \rangle$ (see Fig. [1] panel (A)) but also in the case of the average logarithm
of the term-frequency, $\langle \ln(tf) \rangle$, as displayed in Fig. S7. The computed parameters of the maximum entropy model associated to the discrete $tf$, i.e. a powerlaw with cutoff, outline the trend followed by the maximum entropy distribution of the concepts. We recall from what we discussed in the main text that the exponent of the powerlaw part, $s$, is characteristic of the process that drive the observed distribution of the $tf$. The striking maximum shown by its distribution in Fig. 2 is particularly significant: it is at $s = 3/2$, a value which strongly suggests the presence of a branching process à la Galton-Watson type behind distribution which gives rise to the observed $tf$ sequence. For the sake of completeness, we display the distribution of the exponential cutoff $\lambda$ in Fig. S8. Note that, in this case, the distribution is peaked around 0.

![Graph A](image1)

**FIG. S5.** Distribution of concept features. Panel (A) shows the distribution of the document frequency, $df$, while panel (B) refers to the distribution of the average term-frequency, $\langle tf \rangle$. In both panels, we reported the powerlaw fit of the distribution along with the powerlaw parameter $\alpha$ and its standard deviation. Powerlaw fits are displayed to highlight the trend of the distributions but are not meant to be the best fitting models. Nevertheless, both distributions have a broad shape, indicating that the variability of their values is high, eventually lacking a scale that sets a typical value.

![Graph B](image2)

**FIG. S6.** Two-dimensional tessellation filtering for the Physics 2013 dataset. In panel (A) the shape and color of the points account for the region they fall into (see Fig. S1). Ubiquitous concepts (red starts) are 4.6 %, rare/specific concepts (cyan triangles) are 11 %, significant concepts (green circles) are 46 %, while the remaining concepts (orange squares) are 39 %. Black diamonds represent concepts marked as generic in the ScienceWISE ontology and the solid black line is a guideline of their average position. In panel (B), instead, points are colored by the percentile $p$ of residual entropy they belong to, as reported in the color bar. This figure is the same displayed in the bottom right corner of panel (C) in Fig. 1. Logarithmic scales are used to improve the visualization.
FIG. S7. Relation between the conditional entropy $S_c$ and the average logarithm of the term-frequency, $\langle \ln(tf) \rangle$. Black diamonds represent SW common concepts (CC) which are scattered in a very tiny region at the boundary of the plot.

FIG. S8. Distribution of the exponential cutoff parameter, $\lambda$, of the maximum entropy model for the discrete $tf$ case in the Physics 2013 collection. We observe a concentration of the values around zero.

3. Differences between $S_d$ and IDF rankings

In Fig. S9 we plot the overlap score $O$, (Eq. S20), of the lists of concepts belonging respectively to the $n$-th and $m$-th percentile slices of $IDF$ and $S_d$. In the case of $IDF$, concepts are ranked from the most frequent (i.e. having the smallest $IDF$) to the least one. In the case of residual entropy, instead, we rank concepts from the closest to its maximum entropy (i.e. smallest $S_d$) to the further away. According to the definition of $O$, matrices are normalized by row. The analysis of the overlap matrix denotes the presence of a certain degree of similarity in the region near the main diagonal. Within such region, with the sole exception of $O_{10,10}$, the average overlap is around 15% indicating that – in general – more frequent concepts tend to fall in higher percentiles of the residual entropy. The $O_{10,10}$ element, instead, has a value around 50%, denoting a remarkable affinity between these sets. This means that generic concepts are, to some extent, also those appearing more often within the collection albeit this is not
always the case.

FIG. S9. Overlap between lists of concepts ranked according to the residual entropy $S_d$ and $IDF$ for the Physics 2013 collection. The color of the cells denote the amount of overlap $\mathcal{O}$. Matrices are normalized by row and white entries correspond to absence of overlap. The dashed line indicates the first diagonal.

4. Comparison between communities of networks obtained with different filtering criteria

In Fig. S10 we report the heatmaps of the Jaccard score $J$ (see Sec. SI E 2) computed between the communities of the similarity networks obtained by pruning out a given fraction $f$ of concepts according either to their $IDF$ ranking (horizontal axis) or to their $S_d$ one (vertical axis). Each column corresponds to a different amount of removed concepts spanning from 10% to 30%. Overall, the Jaccard heatmaps tell us that there is always a certain degree of similarity among the communities found after filtering according to $IDF$ and $S_d$. However, the overlap fades away as the system begins to display a richer topic/community organization in response to the increasing amount of concepts removed. More specifically, as $p$ increases, we notice the coexistence of one group of communities present in both networks and another group of communities characteristic of a given filtering criterion. Such coexistence is yet another proof (as already shown in Secs. SI B, SI E 1 and SII A 3) that using residual entropy to filter the network is not equivalent to the filtering based on the other previously existing methods.

5. Ranking of concepts within papers

In the following, we want to understand if the entropic selection of concepts could be used also to rank concepts and, in turns, use those rankings to classify documents. To this aim, we select ten highly cited documents from the collection (see Tab. Siii), and for each of them we consider rankings based on TF, $IDF$, TF-IDF and $S_d$ respectively. Next, we order concepts from the most generic to the least one according to the four rankings. This translates into sorting either in descending order (TF, TF-IDF) or in ascending one ($IDF$, $S_d$). The results are reported in Tab. Siv.

Qualitatively speaking, the ranked lists of concepts presented in Tab. Siv seems to confirm, on one hand, the inability of $S_d$ and $IDF$ to fully grasp the subject of each article. On the other hand, instead, TF and TF-IDF perform remarkably better in the same task. However, these conclusions are not surprising: both $S_d$ and $IDF$ are global measures, defined on the whole collection in order to quantify the importance of a concept for the entire collection. Conversely, both TF and TF-IDF are local measures that capture the significance of a concept within papers.
FIG. S10. Overlap among communities of the similarity network obtained filtering concepts either using entropy (y-axis) or IDF (x-axis). The color of the cells denotes the Jaccard score $J$. Each label accounts for the main topic and the size of a give community. Each matrix refers to removing, respectively, the 10% (A), 20% (B) and 30% (C) of the concepts. The values of the overlap in panel (A) range from 0.70 to 0.45 on the main diagonal, while off-diagonal elements are below 0.11, except for the overlap between “ghost” communities which is 0.34. Panel (B) features overlaps between 0.59 and 0.22 on the main diagonal, while the other values are below 0.17 and the overlap between “ghost” communities is 0.29. Finally, panel (C) displays values ranging from 0.54 to 0.19 on the main diagonal and below 0.18 outside, apart from the “ghost” communities overlap which is 0.30.

TABLE Siii. Main attributes of the manuscripts selected to study the rankings of concepts within documents. For each document we report its arXiv ID, the number of citations, #cit, its principal category and eventual secondary ones. Finally, we provide the publication venue.

| arXiv ID        | #cit | Princ. cat.         | Other cat(s)         | Venue                                   |
|-----------------|------|---------------------|----------------------|-----------------------------------------|
| 1306.5856       | 1103 | cond-mat.mtrl-sci   | –                    | Nat Nanotech 8, 235–246 (2013)          |
| 1301.0842       | 310  | astro-ph.EP         | –                    | Astroph Jour 766, 81 (2013)             |
| 1308.0321       | 365  | cond-mat.quant-gas  | cond-mat.str-el, quant-ph | Phys Rev Lett 111, 185301 (2013)     |
| 1301.1340       | 223  | hep-ph              | –                    | Rep Prog Phys 76, 056201 (2013)        |
| 1301.0319       | 261  | astro-ph.SR         | astro-ph.IM          | Astrophys J Suppl Ser 208, 4 (2013)    |
| 1304.6875       | 499  | astro-ph.HE         | astro-ph.SR, cond-mat.quant-gas, gr-qc | Science 340, Issue 6131 (2013)  |
| 1306.2314       | 170  | astro-ph.CO         | –                    | Phys Rev D 88, 043502 (2013)          |
| 1311.6806       | 169  | astro-ph.EP         | –                    | PNAS 110, 19175 (2013)                |
| 1302.5433       | 150  | cond-mat.supr-con   | cond-mat.mes-hall    | J Phys Conden. Matter 25, 233201 (2013) |
| 1303.3572       | 22   | cond-mat.str-el     | hep-ph, quanth-ph   | Phys Rev B 89, 045127 (2014)          |

TABLE Siv: List of the ten most generic concepts for the papers listed in Tab. Siii. We rank concepts using: residual entropy $S_d$, Inverse Document Frequency IDF, term frequency TF and TF-IDF. Concepts indicated as common by ScienceWISE, are marked by an asterisk. The column corresponding to the best ranking is highlighted.

| arXiv ID        | $S_d$ | IDF   | TF    | TF-IDF |
|-----------------|-------|-------|-------|--------|
| 1306.5856       |       |       |       |        |
| Raman spectroscopy as a versatile tool for studying the properties of graphene | Experimental data * | Energy * | Measurement * | Phonon * | Phonon |
|                  | Regularization | Intensity | Regularization | Measurement | measurement |
|                  | Temperature * | Field * | Potential * | Mass * | Raman spectroscopy |
|                  | Field * | Optics | Particles * | Probability * | Probability * | Raman scattering |
|                  | Electromagnet * | Electromagnet * | Electromagnet * | Energy * | Energy * | Electron * |
|                  | Energy * | Energy | Temperature * | Energy * | Energy * | Carbonate * |

Continued on next page
continued from previous page

| arXiv ID | $S_d$ | IDF | TF | TF-IDF |
|----------|-------|-----|----|--------|
| 1301.0842 | The false positive rate of Kepler and the occurrence of planets |
| Order of magnitude * | Measurement * | Planet | Planet |
| Numerical simulation | Field * | Star | Kepler Objects of Interest |
| Space telescopes | Potential * | Kepler Objects of Interest |
| Temperature * | Mass * | Period * | False positive rate |
| Statistical error | Temperature * | Frequency * | Star |
| Field * | Probability * | Signal to noise ratio | Eclipsing binary |
| Optics * | Units * | False positive rate | Eclipses |
| Mass * | Frequency * | Eclipsing binary | Signal to noise ratio |
| Frequency * | Periodate * | Eclipses | Neptune |
| Fluctuation * | Velocity * | Orbit * | Earth-like planet |
| Intensity | Selection rule | Stellar classification |

| 1308.0321 | Realization of the Hofstadter Hamiltonian with ultracold atoms in optical lattices |
| Experimental data * | Energy * | Atom * | Atom * |
| Intensity | Measurement * | Magnetic field * | Magnetic field * |
| Strong interactions | Field * | Potential * | Optical lattice |
| Field * | Potential * | Measurement * | Cyclotron |
| Optics * | Mass * | Optical lattice | Ultracold atom |
| Energy * | Particles * | Hamiltonian | Spin Quantum Hall Effect |
| Mass * | Units * | Spin * | Band mapping |
| Wavelength * | Electron * | Orbit * | Time-reversal symmetry |
| Frequency * | Frequency * | Cyclotron | Hamiltonian |
| Factorization | Periodate * | Energy * | Superlattice |

| 1301.1340 | Neutrino Mass and Mixing with Discrete Symmetry |
| Order of magnitude * | Energy * | Symmetry * | Neutrino mass |
| Experimental data * | Measurement * | Neutrino mass | Neutrino |
| Weak interaction | Field * | Mass * | Charged lepton |
| Vacuum * | Potential * | Neutrino | Symmetry * |
| Field * | Mass * | Charged lepton | Sterile neutrino |
| Electromagnet * | Particles * | Field * | See-saw |
| Energy * | Probability * | Leptons * | Leptons * |
| Mass * | Units * | Sterile neutrino | Mixing patterns |
| Equation of motion * | Vector * | Vacuum expectation value | Grand unification theory |
| Momentum * | Electron * | See-saw | Vacuum expectation value |

| 1301.0319 | Modules for Experiments in Stellar Astrophysics (MESA): Giant Planets, Oscillations, Rotation, and Massive Stars |
| Order of magnitude * | Energy * | Mass * | Star |
| Stellar physics | Measurement * | Star | White dwarf |
| Right Hand Side | Field * | Frequency * | Massive stars |
| of the expression * | Potential * | White dwarf | Stellar evolution |
| Regularization | Mass * | Angular momentum * | Angular momentum * |
| Intensity | Particles * | Massive stars | Mass * |
| Temperature * | Temperature * | Temperature * | Planet |
| Field * | Probability * | Pressure * | Red supergiant |
| Optics * | Units * | Luminosity | Asteroseismology |
| Energy * | Electron * | Stellar evolution | Zero-age main |
| Mass * | | | sequence stars |

| 1304.6875 | A Massive Pulsar in a Compact Relativistic Binary |
| Order of magnitude * | Energy * | Mass * | White dwarf |
| Stellar physics | Measurement * | White dwarf | Neutron star |
| Solar mass | Field * | Orbit * | Pulsar |
| Temperature * | Potential * | Neutron star | Orbit * |
| Statistical error | Mass * | Pulsar | Gravitational wave |
| Degree of freedom | Particles * | General relativity | General relativity |
| Field * | Temperature * | Gravitational wave | Companion |
| Optics * | Probability * | Star | Mass * |
| Energy * | Units * | Gravitation * | Low-mass X-ray binary |

Continued on next page
| arXiv ID     | $S_d$ | IDF | TF | TF-IDF |
|--------------|-------|-----|----|--------|
| **Mass**     | Vector | Companion | Binary star |
| 1306.2314    | 1306.2314 | Warm Dark Matter as a solution to the small scale crisis: new constraints from high redshift Lyman-alpha forest data |
| Astrophysics and cosmology | Measurement * | Simulations * | WDM particles |
| Numerical simulation | Mass * | Resolution * | Simulations * |
| Regularization | Particles * | Cold dark matter | Cold dark matter |
| Intensity | Temperature * | Temperature * | Intergalactic medium |
| Temperature * | Probability * | Intergalactic medium | Mean transmitted flux |
| Statistical error | Universe * | Quasar | Ultraviolet background |
| Degree of freedom | Velocity * | WDM particles | Quasar |
| Optics * | Objective * | Wavenumber * | Free streaming |
| Mass * | Formate * | Free streaming | Redshift bins |
| Fluctuation * | Optics * | Matter power spectrum | Matter power spectrum |

| 1311.6806 | Prevalence of Earth-size planets orbiting Sun-like stars |
| Stefan-Boltzmann constant | Energy * | Signal to noise ratio | Kepler Objects of Interest |
| Solar mass | Measurement * | Light curve | Signal to noise ratio |
| Intensity | Potential * | Photometry | Light curve |
| Temperature * | Mass * | Habitable zone | Photometry |
| Statistical error | Temperature * | Stellar radii | Stellar radii |
| Energy * | Probability * | Extrasolar planet | Extrasolar planet |
| Mass * | Periodate * | Eclipsing binary | Eclipsing binary |
| Wavelength * | Universe * | Habitable zone | High resolution |
| Fluctuation * | Objective * | Eclipsing binary | High resolution |
| Uniform distribution | Statistics | Event * | High resolution |
| **Statistical error** | **Probability** | **Temperature** | **Eclipsing binary** |
| **Fluctuation** | **Objective** | **Temperature** | **Event** |
| **Uniform distribution** | **Statistics** | **Event** | **High resolution** |

| 1302.5433 | Majorana Fermions in Semiconductor Nanowires: Fundamentals, Modeling, and Experiment |
| Order of magnitude * | Energy * | Majorana fermion | Majorana fermion |
| Bohr magneton | Measurement * | Energy * | Nanowire |
| Experimental data * | Field * | Nanowire | Majorana bound state |
| of the exression * | Potential * | Superconductor | Superconductor |
| Critical value | Mass * | Topology * | Semiconductor |
| Regularization | Particles * | Field * | Josephson effect |
| Temperature * | Temperature * | Semiconductor | Topology * |
| Expectation Value | Probability * | Hamiltonian | Superconductivity |
| Degree of freedom | Units * | Superconductivity | Topological superconductor |
| Field * | Vector * | Measurement * | Heterostructure |

| 1303.3572 | 3-dimensional bosonic topological insulators and its exotic electromagnetic response |
| Right Hand Side | Energy * | Bosonization | Dyon |
| of the exression * | Field * | Dyon | Electromagnetism |
| Regularization | Field * | Charge * | U(1) * |
| Strong interactions | Potential * | Condensation | Witten effect |
| Degree of freedom | Mass * | Fermion * | Bosonization |
| Vacuum * | Particles * | Statistics | Condensation |
| Field * | Units * | Symmetry * | Projective construction |
| Electromagnet * | Vector * | Electromagnetism | Time-reversal symmetry |
| Energy * | Periodate * | Symmetry * | Fermion * |
| Mass * | Symmetry * | Time-reversal symmetry | Mean field |

Tab. Siv shows how the generality of a concept depends on the criterion used to rank it. It is also worth to see how the selective removal of concepts reverberates on the rankings. To this aim, we report in Tab. Sv the ten most generic concepts as a function of the filtering intensity $p$ going from the original set ($p = 0\%$) to the optimal level ($p_{opt} = 30\%$) as defined in Sec. SIII A. At first glance, we observe how increasing the aggressiveness of the filter produces an immediate decrease of the number of concepts marked as common by SW. More importantly, we clearly see how the pruning removes also concepts classifiable as generic that have not been marked as such by SW.
TABLE Sv: List of the ten most generic concepts per paper as a function of the entropic filtering intensity $p$. $p = 0$ denotes the original set, while $p_{opt}$ corresponds to the optimal level of filtering. Concepts indicated as common by ScienceWISE are marked by an asterisk.

| arXiv ID | $p = 0$ | $p = 10$ | $p = 20$ | $p_{opt} = 30$ |
|----------|---------|----------|----------|----------------|
| 1306.5856 | Raman spectroscopy as a versatile tool for studying the properties of graphene | Electron transition | Electron hole pair | Monochromator |
|          | Regularization | Irradiance | Topological insulator | Surface plasmon resonance |
|          | Intensity | Group velocity * | Thermal Expansion | Bilayer graphene |
|          | Temperature | Reciprocal lattice | Transistors | Graphene layer |
| Field * | Field * | Diffraction * | Backscattering | Superlattice |
| Optics * | Optics * | Nanostructure | Scanning tunneling microscope | Van Hove singularity |
| Electromagnet * | Electromagnet * | Hydrostatics | Graphite | Surface plasmon |
| Energy * | Energy | Electron scattering | Nitriding | Exciton |
| Mass * | Mass * | Circular polarization * | Dirac point | Nanomaterials |
| Wavelength * | Wavelength * | Space-time singularity | Normal mode | Intervalle scattering |

1301.0842 The false positive rate of Kepler and the occurrence of planets

| Order of magnitude * | Planet formation | Stellar classification | Luminosity class |
| Numerical simulation | Near-infrared | Early-type star | Eclipsing binary |
| Space telescopes | Error function | Probability density function * | Matched filter |
| Temperature * | Companion | Companion stars | Astroseismology |
| Statistical error | Spectrographs | Star counts | High accuracy radial velocity |
| Field * | Angular distance | Earth-like planet | planetary search |
| Optics * | Stellar magnitude | Orbital elements | Hot Jupiter |
| Mass * | Extinction | Eccentricity | Neptune |
| Frequency * | Kurtsomogov-Smirnov test | Primary stars | Periastron |
| Fluctuation * | Solar neighborhood | Giant planet | Planet |

1308.0321 Realization of the Hofstadter Hamiltonian with ultracold atoms in optical lattices

| Experimental data * | Atomic number * | Coriolis force * | Landau-Zener transition |
| Intensity | Helicity | Topological insulator | Chern number |
| Strong interactions | Quantum Hall Effect | Mott insulator | Superlattice |
| Field * | SU(2) * | Cyclotron | Magnetic trap |
| Optics * | Freezing | Edge excitations | Spin Hall effect |
| Energy * | Lorentz force * | Topological order | Spin Quantum Hall Effect |
| Mass * | Spontaneous emission | Berry phase | Quadrupole magnet |
| Wavelength * | Optical lattice | Fractal | Band mapping |
| Frequency * | Quadrupole | Landau-Zener transition | Lowest Landau Level |
| Factorisation | Bose-Einstein condensate | Chern number | Hofstadter’s butterfly |

1301.1340 Neutrino Mass and Mixing with Discrete Symmetry

| Order of magnitude * | Neutron * | Zenith | Flavour physics |
| Experimental data * | Antisymmetrizer | Supersymmetry breaking | Atmospheric neutrino |
| Weak interaction | Mass spectrum | Upper atmosphere | Infinite group |
| Vacuum * | Supersymmetry | Weak neutral current | Clebsch-Gordan |
| Field * | Baryon number | interaction | coefficients |
| Electromagnet * | Subgroup | CP violation | Neutrino telescope |
| Energy * | Permutation | Euler angles | |
| Mass * | Quark mass | Rotation group * | CP violating phase |
| Equation of motion * | Irreducible representation | Superpotential | Proton decay |
| Momentum * | Embedding | Reactor neutrino experiments | Neutrino mixing angle |
| купить * | | | Complex conjugate |
| | | | representation |
| | | | | Neutralino |

1501.0319 Modules for Experiments in Stellar Astrophysics (MESA): Giant Planets, Oscillations, Rotation, and Massive Stars

| Order of magnitude * | Planet formation | Diffusion equation | Complete mixing |
| Stellar physics | Accretion | Gravitational energy | Kelvin-Helmholtz timescale |
| Right Hand Side | Low-mass stars | Circumstellar envelope | Radiative Diffusion |
| of the expression * | | | Optical bursts |
| Regularization | Massive stars | Early-type star | Asteroseismology |
| Intensity | Diffusion coefficient | Helium shell flashes | |
### 1304.6875  A Massive Pulsar in a Compact Relativistic Binary

| Order of magnitude * | Accretion | Radio telescope | Stellar oscillations |
|----------------------|-----------|----------------|----------------------|
| Stellar physics      | Massive stars | Moment of inertia * | Mass discrepancy |
| Solar mass           | Black hole | Mass function | Matched filter |
| Temperature *        | Irradiance | Comparison stars | Zero-age main sequence stars |
| Statistical error    | Sloan Digital Sky Survey | Circumstellar envelope | Grism |
| Degree of freedom    | Cooling | Probability density function * | Radio pulsar |
| Field *              | Companion | Companion stars | Laser Interferometer |
| Optics *             | Spectrographs | Roche Lobe | Gravitational-Wave Observatory |
| Energy *             | Space-time singularity | Mass accretion rate | Radiation damping |
| Mass *               | Stellar surfaces | Peculiar velocity | Barycenter |
| Fluctuation *        | - | - | Space velocity |

### 1306.2314  Warm Dark Matter as a solution to the small scale crisis: new constraints from high redshift Lyman-alpha forest data

| Astrophysics and cosmology * | Simulations * | Dark matter particle | Nuisance parameter |
|-------------------------------|---------------|---------------------|--------------------|
| Numerical simulation          | -             | -                   | Satellite galaxy   |
| Regularization                | Mean transmitted flux | Matter power spectrum | Free streaming     |
| Intensity                     | Sloan Digital Sky Survey | Luminosity function | Quasar             |
| Temperature *                 | Cooling | A dwarfs | Active Galactic Nuclei |
| Statistical error             | Spectrographs | Supernova | Planck data |
| Degree of freedom             | Dark matter | Tellurite | Halo finding algorithms |
| Optics *                      | Wavenumber | Monte Carlo Markov chain | Baryon acoustic oscillations |
| Energy *                      | Absorption feature | Cold dark matter | Void |
| Mass *                        | -             | -                   | Strong gravitational lensing |
| Fluctuation *                 | -             | -                   | - |

### 1311.6806  Prevalence of Earth-size planets orbiting Sun-like stars

| Stefan-Boltzmann constant | Simulations * | Hertzsprung-Russell diagram | Earth-like planet |
|---------------------------|---------------|-----------------------------|-------------------|
| Solar mass                | Companion | Monte Carlo Markov chain | Limb darkening |
| Intensity                 | Stellar surfaces | Hydrogen 21 cm line | Planet |
| Temperature *             | Stellar magnitude | Keck Array | High resolution |
| Statistical error         | Host star | Parallax | échelle spectrometer |
| Energy *                  | Orbit Eccentricity | Eclipsing binary | Eclipses |
| Mass *                    | Droplet | Habitability | Habitat zone |
| Wavelength *              | Angular separation | Asteroseismology | Gaussian process |
| Fluctuation *             | Teams | Limb darkening | Mars |
| Uniform distribution      | Galactic structure | Planet | Orange dwarf |

### 1302.5433  Majorana Fermions in Semiconductor Nanowires: Fundamentals, Modeling, and Experiment

| Order of magnitude * | Tight-binding model | Second quantization | P-wave |
|----------------------|---------------------|---------------------|--------|
| Bohr magneton        | Quantum dots | Feshbach resonance | Quantum decoherence |
| Experimental data *  | Neutron *           | Zero mode | Nanowire |
| Right Hand Side      | Rest mass *         | Topological insulator | Chern number |
| of the exression *   | -                   | -                   | Topological superconductor |
| Critical value       | Nanostructure | Proximity effect | Local density of states |
| Regularization       | Winding number | Networks * | Topological superconductor |
| Temperature *        | Helicity | Critical current | Andreev reflection |
| Expectation Value    | Quantum Hall Effect | Scaling limit | Josephson effect |
| Degree of freedom    | Coarse graining | Pair potential | Weak antilocalization |
| Field *              | Chiral symmetry | Quantum critical point | Non-Abelian statistics |
The information presented in Tab. Sv confirms the power of our filtering methodology. In analogy to what we have done in Tab. Siv, we check if $S_d$ still outperforms other rankings also in the filtered networks. For this reason, in Tab. Svi we report the rankings of the concepts at the optimal level of filtering ($p_{opt} = 30\%$). A quick glance at its columns tells us that, albeit being more specific, concepts ranked using $S_d$ are still capable of describing the content of the document, outperforming measures like TF and TF-IDF.

TABLE Svi: Ten most generic concepts ranked upon different indices among the set of concepts available at the $S_d$ optimal level of filtering ($p_{opt} = 30\%$). Columns are the same as Tab. Siv. We highlight the columns of $S_d$ (best ranking) and TF-IDF (standard ranking).

| arXiv ID | $S_d(p_{opt} = 30)$ | IDF ($p_{opt} = 30$) | TF ($p_{opt} = 30$) | TF-IDF ($p_{opt} = 30$) |
|----------|---------------------|----------------------|--------------------|-------------------------|
| 1306.5856 | Raman spectroscopy as a versatile tool for studying the properties of graphene | Exciton | Surface enhanced | Surface enhanced |
|          | Monochromator | Surface plasmon resonance | Superlattice | Raman spectroscopy |
|          | Bilayer graphene | Graphene layer | Van Hove singularity | Kohn anomaly |
|          | Graphene layer | Bilayer graphene | Gr"uneisen parameter | Gr"uneisen parameter |
|          | Superlattice | Surface plasmon | Hexagonal boron nitride | Van Hove singularity |
|          | Van Hove singularity | Monochromator | Nanocrystalline | Nanocrystalline |
|          | Surface plasmon | Van Hove singularity | Graphene layer | Graphene layer |
|          | Exciton | Nanocrystal | Exciton | Fullerene |
|          | Nanomaterials | S-process | Nanocrystal | Nanocrystal |
|          | Intervalance scattering | Fullerene | Fullerene | Depolarization ratio |

1301.0842 The false positive rate of Kepler and the occurrence of planets

| Luminosity class | Planet | Planet | Planet |
|------------------|--------|--------|--------|
| Eclipsing binary | White dwarf | Kepler Objects of Interest | Kepler Objects of Interest |
| Matched filter | Eclipses | False positive rate | False positive rate |
| Asteroseismology | M dwarfs | Eclipses | Eclipsing binary |
| High accuracy radial velocity planetary search | Periastron | Eclipsing binary | Eclipses |
| Hot Jupiter | Eclipsing binary | Neptune | Neptune |
| Triple system | Hot Jupiter | Super-earth | Super-earth |
| Periastron | Neptune | Triple system | Triple system |
| Planet | Asteroseismology | White dwarf | Logarithmic distribution |
|         | Super-earth | Logarithmic distribution | White dwarf |

1308.0321 Realization of the Hofstadter Hamiltonian with ultracold atoms in optical lattices

| Landau-Zener transition | Superlattice | Spin Quantum Hall Effect | Spin Quantum Hall Effect |
|-------------------------|--------------|--------------------------|--------------------------|
| Chern number            | Chern number | Superlattice             | Band mapping             |
| arXiv ID | $S_d(p_{opt} = 30)$ | IDF ($p_{opt} = 30$) | TF ($p_{opt} = 30$) | TF-IDF ($p_{opt} = 30$) |
|---------|---------------------|----------------------|---------------------|-------------------------|
| Superlattice | Spin Hall effect | Band mapping | Superlattice |
| Magnetic trap | Lowest Landau Level | Landau-Zener transition | Landau-Zener transition |
| Spin Hall effect | Spin Quantum Hall Effect | Spin Hall effect | Spin Hall effect |
| Spin Quantum Hall Effect | Magnetic trap | Magnetic trap | Quadrupole magnet |
| Quadrupole magnet | Landau-Zener transition | Hofstadter’s butterfly | Hofstadter’s butterfly |
| Band mapping | Hofstadter’s butterfly | Chern number | Magnetic trap |
| Lowest Landau Level | Quadrupole magnet | Lowest Landau Level | Lowest Landau Level |
| Hofstadter’s butterfly | Band mapping | Quadrupole magnet | Chern number |

1301.1340 Neutrino Mass and Mixing with Discrete Symmetry

| Flavour physics | Gamma ray burst | Mixing patterns | Mixing patterns |
| Atmospheric neutrino | Superfield | Solar neutrino | Tri Bimaximal mixing |
| Infinite group | Neutralino | Reactor Experiment | Solar neutrino |
| Clebsch-Gordan coefficients | Mantle | Tri Bimaximal mixing | Reactor Experiment |
| Neutrino telescope | Two Higgs Doublet Model | Clebsch-Gordan coefficients | Neutrino Oscillation |
| CP violating phase | Supermultiplet | Super-Kamiokande | Neutrino Oscillation |
| Proton decay | CP violating phase | SNO+ | Super-Kamiokande |
| Neutrino mixing angle | Atmospheric neutrino | Atmospheric neutrino | Clebsch-Gordan coefficients |
| Complex conjugate representation | Proton decay | Trimaximal mixing | Mikhail-Smirnovov- |
| Neutrino | Massive neutrino | Type I seesaw | Wolfenstein effect |

1301.0319 Modules for Experiments in Stellar Astrophysics (MESA): Giant Planets, Oscillations, Rotation, and Massive Stars

| Complete mixing | Planet | White dwarf | White dwarf |
| Kelvin-Helmholtz timescale | White dwarf | Planet | Planet |
| Radiative Diffusion | Gamma ray burst | Zero-age main sequence stars | Red supergiant |
| Optical bursts | Optical bursts | Red supergiant | Astroseismology |
| Asteroseismology | Pre-main-sequence star | Pre-main-sequence star | Pre-main-sequence star |
| Stellar oscillations | Asymptotic giant branch | Pre-main-sequence star | Stellar oscillations |
| Zero-age main sequence stars | Zero-age main | Asymptotic giant branch | Stellar oscillations |
| Classical nova | Large Synoptic Survey Telescope | Gamma ray burst | Asymptotic giant branch |
| Large Synoptic Survey Telescope | Wolf-Rayet star | Stellar oscillations | Gamma ray burst |
| Giant branches | Asteroseismology | Optical bursts | Classical nova |

1304.6875 A Massive Pulsar in a Compact Relativistic Binary

| Lunar Laser Ranging experiment | Planet | White dwarf | White dwarf |
| Mass discrepancy | Pulsar | Pulsar | Pulsar |
| Matched filter | White dwarf | Low-mass X-ray binary | Low-mass X-ray binary |
| Zero-age main sequence stars | Albedo | Binary pulsar | Binary pulsar |
| Grism | VLT telescope | Orbital angular momentum of light | Green Bank Telescope |
| Radio pulsar | Low-mass X-ray binary | Green Bank Telescope | Orbital angular momentum of light |
| Laser Interferometer Observatory | Zero-age main sequence stars | Zero-age main sequence stars | Zero-age main sequence stars |
| Radiation damping | Grism | Millisecond pulsar | Solar system barycenter |
| Barycenter | Laser Interferometer Observatory | Solar system barycenter | Radio pulsar |
| Space velocity | Millisecond pulsar | VLT telescope | Dispersion measure |

Continued on next page
### 1306.2314

**Warm Dark Matter as a solution to the small scale crisis: new constraints from high redshift Lyman-alpha forest data**

| Nuisance parameter | Active Galactic Nuclei | Quasar | WDM particles |
|--------------------|------------------------|--------|---------------|
| Satellite galaxy   | Quasar                 |        | Ultraviolet background |
| Free streaming     | Gamma ray burst        |        | Free streaming |
| Quasar             | Void                   |        | Ultraviolet background |
| Active Galactic Nuclei | Baryon acoustic oscillations | Reionization | Redshift bins |
| Planck data        | Reionization           |        | Temperature-density relation |
| Halo finding algorithms | Satellite galaxy | Reionization | Warm dark matter |
| Baryon acoustic oscillations | Nuisance parameter | Population III | Effective optical depth |
| Void               | Population III         |        | Reionization   |
| Strong gravitational lensing | Free streaming | Nuisance parameter | WDM particle mass |

### 1311.6806

**Prevalence of Earth-size planets orbiting Sun-like stars**

| Eclipsing binary | Planet            | Kepler Objects of Interest | Kepler Objects of Interest |
|------------------|-------------------|----------------------------|----------------------------|
| Asteroseismology | Eclipses          | Eclipses                   | Habitable zone             |
| Limb darkening   | Eclipsing binary  | Eclipsing binary           | High resolution            |
| Planet           | Mars               | Mars                        | Mars                       |
| High resolution  | Habitable zone     | High resolution             | Habitable zone             |
| échelle spectrometer | Habitable zone | Habitable zone             | Habitable zone             |
| Eclipses         | Habitable zone     | Habitable zone             | Habitable zone             |
| Gaussian process | Gaussian process   | Limb darkening              | Limb darkening             |
| Mars             | Ephemerides        | False positive rate         | False positive rate        |
| Orange dwarf     | High resolution   | Ephemerides                | Orange dwarf               |
| échelle spectrometer | High resolution | Ephemerides                | Orange dwarf               |

### 1302.5433

**Majorana Fermions in Semiconductor Nanowires: Fundamentals, Modeling, and Experiment**

| P-wave | Nanowire | Nanowire | Nanowire |
|--------|----------|----------|----------|
| Quantum decoherence | Carbon nanotubes | Majorana bound state | Majorana bound state |
| Nanowire | Local density of states | Josephson effect | Josephson effect |
| Chern number | Local density of states | Topological superconductor | Topological superconductor |
| Local density of states | Chern number | P-wave | P-wave |
| Topological superconductor | Topological superconductor | Local density of states | Local density of states |
| Andreev reflection | Andreev reflection | Fermion doubling | Fermion doubling |
| Josephson effect | Josephson effect | Andreev reflection | Andreev reflection |
| Weak antilocalization | Weyl fermion | Non-Abelian statistics | Non-Abelian statistics |
| Non-Abelian statistics | Non-Abelian statistics | Moore-Read | Moore-Read |
|                     |                      | Pfaffian wavefunction | Pfaffian wavefunction |

### 1303.3572

**3-dimensional bosonic topological insulators and its exotic electromagnetic response**

| Hall conductance | Exciton | Dyon | Dyon |
|------------------|---------|------|------|
| Electric magnetic | Axion   | Witten effect | Witten effect |
| Long-range entanglement | Hall conductance | Projective construction | Group cohomology |
| Topological field theory | Topological field theory | Topological field theory | Topological field theory |
| Axion            | Electric magnetic | Exciton | Response theory |
| Exciton          | Long-range entanglement | Symmetry protected topological order | Charge quantization |
| Short-range entanglement | Symmetry protected topological order | Charge quantization | Short-range entanglement |
| Symmetry protected topological order | Dyon | Axion | Symmetry protected topological order |
| Group cohomology | Short-range entanglement | Axion | Long-range entanglement |
| Charge quantization | Charge quantization | Hall conductance | Hall conductance |
B. Climate change web documents

In this section, we present the results obtained for another dataset originated from a collection of web documents on climate change. It is worth stressing that, on average, a web document lacks the same structural organization of a scientific manuscript. This is partially due to the fact that such kind of documents convey information in a perspective different than those of a scientific document on the same topic. Web documents have almost no limitations in length, they can simply report facts (news releases) or provide an opinion on a given subject. Moreover, contrary to the case of Physics, the ScienceWISE platform does not have an ontology on climate change. Therefore, we do not have access to the validated concepts but, instead, directly to the keywords extracted by ScienceWISE machinery.

1. Source

Our collection of web documents has been extracted from the pool of documents available within the ScienceWISE database. More specifically, SW has a collection of tweets on climate change, posted between January and June 2015, that was compiled using Twitter API [59] through several harvesting campaigns. From the original pool of tweets, we kept only the original ones, thus culling mentions, re-tweets and other similar “non original” posts. From the original tweets, we kept only those written in English containing at least one URL. Such procedure returns a list of distinct URLs pointing to web documents (approx. 50 millions) ranked according to the number of tweets pointing to a each URL. We then select the top 100,000 most “tweeted” URLs. To ensure tematic consistency with the “climate change” topic, only documents with at least one of 165 specific concepts in their URL are retained.

The above procedure returns a set of 30705 documents. To discard documents excessively short, we decided to consider only those with, at least, $L_{\text{min}} = 500$ words. After this thresholding, our collection is made of 18770 documents. To extract the keywords from these documents (and the tweets associated with them) we used the KPEX algorithm [53], since it is natively implemented in the SW platform. The KPEX extraction returns 822601 unique keywords which were stemmed first and then lemmatized, returning a final set of 152871 keywords. Since SW platform does not have a cured ontology on climate change, n-grams extracted from KPEX are simple keywords and thus cannot be called concepts in a strict sense, albeit we will continue to do so in the rest of the manuscript. The main features of the similarity network obtained from this collection are reported in the first row of Tab. S.7

2. Entropic filtering

Given the high heterogeneity of the collection in terms of document length $L$, already pointed out in Fig. S.3, we decided to use the $tf$ density, $rtf$, instead of its raw value ending in a lognormal maximum entropy distribution (see Sec. SI C 2 for details). Moreover, keywords for which $\max(rtf) - \min(rtf) < 0.005$ are ignored, since it is better to discard terms that have similar values of $rtf$. Thus, the number of keywords gets shrunk to 9222. The location of points on the $S_c$, $S_{\text{max}}$ plane is reported in Fig. S.11. As for Physics, we clearly observe a stratification of the residual entropy $S_d$ on the plane confirming the validity of our filtering criterion. Among the generic concepts in the percentile slice $p = 10$ we find “people”, “climate change”, “water”, “home” and “company”. On the other hand, between concepts at the percentile slice $p = 50$ we recognize “palm”, “whale”, “Boulder”, “metal” and “shop”. The selective removal of concepts based on $S_d$ alters the topological properties of the similarity network, causing its overall sparsification as reported in Tab. S.7

3. Sankey Diagram

As for Physics 2013 collection, the Sankey diagram shown in Fig. S.12 is useful to understand the progressive specialization of the organization of topics in response to the selective removal of concepts. However, one significative difference captures our attention: the presence of a remarkable condensation phenomenon taking place around the extreme weather/energy storage communities going from $p = 5\%$ to $p = 20\%$. To gain insight on such phenomenon, for each community, $s$, we study the coverage $\Gamma_s(\tilde{C})$ of the set, $\tilde{C}$, of top twenty most locally used concepts. The coverage of a set of concepts $\Gamma_s(\tilde{C}) \in [0, 1]$ is defined as the union of the sets of documents where those concepts appear divided by the size of number of documents in the community $N^s_d$. Hence, $\Gamma_s(\tilde{C}) = \frac{1}{N^s_d} \bigcup_{c \in \tilde{C}} N^s_a(c)$. The coverage of the community named “???” is $\Gamma = 0.016$ which is pretty small compared to $\Gamma = 0.64$ of “extreme weather” community or $\Gamma = 0.87$ of “ice melting”. The poor coverage of concepts
FIG. S11. Relation between the empirical entropy, $S_c$, and the maximum one, $S_{max}$ for the climate change collection. The colors of the points encode the various percentiles of the residual entropy $S_d$ to which concepts belong to.

![Graph showing relation between $S_c$ and $S_{max}$](image)

\[ 21 \]

TABLE Svi. Topological indicators of the similarity networks between climate change webdocs. The first row ($p = 0\%$) corresponds to the original network, while the others to the networks obtained using the maximum entropy filter. In the columns we report: the percentage of filtered concepts $p$, the number of concepts $N_{con}$, the number of web documents containing at least one concept (nodes) $N_a$, the link density $\rho$, the average $\langle k \rangle$, and maximum degrees $k_{max}$, the average clustering coefficient $\langle C \rangle$, the average path length $\langle l \rangle$ and the number of connected components $M$. The minimum edge weight is equal to $w_{min} = 0.01$.

| $p$ (%) | $N_{con}$ | $N_a$ | $\rho$ (%) | $\langle k \rangle$ | $k_{max}$ | $\langle C \rangle$ | $\langle l \rangle$ | $M$ |
|---------|-----------|-------|------------|-----------------|-----------|------------------|----------------|------|
| 0       | 152871    | 18770 | 10.111     | 1938.624        | 11047     | 0.399            | 1.904          | 1    |
| 5       | 8760      | 18770 | 9.960      | 1869.425        | 10199     | 0.400            | 1.902          | 1    |
| 10      | 8299      | 18762 | 7.610      | 1427.629        | 8677      | 0.480            | 1.936          | 1    |
| 15      | 7838      | 18743 | 5.351      | 1002.891        | 6789      | 0.569            | 2.003          | 1    |
| 20      | 7377      | 18622 | 2.478      | 461.369         | 3863      | 0.658            | 2.221          | 1    |
| 25      | 6916      | 18308 | 0.763      | 139.691         | 1362      | 0.308            | 2.565          | 1    |
| 30      | 6455      | 17888 | 0.512      | 91.521          | 1160      | 0.302            | 2.771          | 1    |
| 40      | 5533      | 16117 | 0.268      | 43.179          | 911       | 0.330            | 3.235          | 14   |
| 50      | 4611      | 13527 | 0.157      | 21.206          | 713       | 0.274            | 3.938          | 43   |
| 60      | 3689      | 10493 | 0.105      | 10.979          | 349       | 0.360            | 5.242          | 147  |
| 70      | 2767      | 7318  | 0.088      | 6.415           | 189       | 0.481            | 8.132          | 443  |
| 80      | 1845      | 4337  | 0.074      | 3.217           | 46        | 0.803            | 10.146         | 925  |
| 90      | 923       | 1876  | 0.102      | 1.919           | 29        | 0.954            | 1.207          | 744  |

in community “???” suggests that documents condense into a single community as a result of similarities associated to small groups of concepts weakly linked together. The intimate origin of such condensation, however, is the presence of keywords whose distribution does not resembles a lognormal, as shown in Fig. S13 for the 10 most frequent ones. Hence, the dissimilarity between the sampled distributions and the lognormal fits is at origin of the feeble ability to describe the content of the webdocs at the community level. Once we get rid of these keywords, the interactions holding together this huge condensed community vanish – or become weaker, at least – and the system breaks apart showing communities that address more specific topics.

4. Differences between $S_d$ and IDF rankings

Using the same formalism of Sec. SII A 3, the overlap between the list of concepts ranked alternatively using $S_d$ or IDF is shown in Fig. S14. The heatmap presents a narrow peak of $O$ located on the main diagonal. Compared with the Physics collection, the overlap is much higher and corresponds exactly to the diagonal elements denoting a much stronger relation between $IDF$ and $S_d$. 
FIG. S12. Static Sankey diagram of the climate change collection. Each community is represented as a colored box whose height is proportional to the number of web documents it contains. A topic is assigned to each box according to the 10 most used keywords, i.e. those appearing in more papers. The thickness of the bands between boxes indicates the number of shared webdocs. Each column denotes a different intensity of filtering \( p \). Concepts are pruned according to their residual entropy \( S_d \) computed from the \( tf \) density, \( rtf \). The minimum fluctuation of \( rtf \) is equal to 0.005. Interactive version available at [50].

FIG. S13. Distribution of the top 10 most frequent concepts within the community of uncertain label "??" found at \( p = 20\% \) in Fig. S12. The lognormal fit of each distribution (see Sec. SIC 2) is plotted with a black line. All the distributions deviate considerably from their lognormal fit, meaning that the Kullback-Leibler distance from the lognormal distribution to the observed one is high.

SIII. FILTERING

In this section we provide some details concerning the choice of the optimal level of filtering (Sec. SIIIA) and how to implement the entropic filtering (Sec. SIIIB).
A. Optimal filtering

In Fig. S15 we report the average size, $\langle N \rangle$, of the communities found at different filtering intensities $p$ together with the size of the ghost community. The optimal value of filtering can be estimated as the maximum of $p$ for which the size of the “ghost” community is below the average size of the communities. In this way, the intersection between these two curves indicates the optimal value of filtering $p_{opt}$ which is equal to 25% in climate change while, instead, in Physics is between 30% and 40% for discrete $tf$ and at 40% for the $tf$ density, $rtf$. 

FIG. S14. Overlap $O$ between lists of concepts ranked according to residual entropy $S_d$ and inverse document frequency $IDF$ for the climate change collection. The matrix is normalized by row and white entries correspond to absence of overlap. The dashed line indicates the main diagonal.

FIG. S15. Average size of communities $\langle N \rangle$ as a function of filter aggressiveness $p$ for both documents collections. Circles refer to the size of the “Ghost” community, while the squares represent the average size of all other communities. The shaded area denotes the standard deviation of the community size.
B. Numerical implementation with pseudocode

In this section we present a step-by-step description of the algorithm used to implement the entropic filtering of concepts. We comment the cases of null model based on power-law distribution with a cutoff first (Sec. SIII B 1) and lognormal then (Sec. SIII B 2). Our pseudocode is written using the Python programming language [60] and we make use of several functions available in the scipy ecosystem [61].

The core of the method is the comparison between two entropies: the actual/experimental one $S_c$ and the expected/theoretical one $S_{max}$ drawn from the the distribution obeying the maximum entropy principle. Given a collection of documents $\mathcal{D}$, for each concept $c$ appearing inside a document $d \in \mathcal{D}$, ScienceWISE provides its boosted term-frequency $tf_d(c)$. Such quantity encodes the relevance of the concept taking into account where it appears. If we consider each document formed by three parts: title, abstract and body, the boosted $tf$ is given by the sum of these contributions:

- the number of times $c$ appears in the body;
- the number of times $c$ appears in the abstract multiplied by a factor of three;
- the number of times $c$ appears in the title, multiplied by a factor of five.

It is worth mentioning that, in the case of $tf$ density, this quantity is divided for the length of document $d$, $L(d)$, hence:

$$ rtfc(d) = \frac{tf_c(d)}{L(d)}. $$

1. Discrete $tf$

Given a sequence of $M$ values $\mathcal{X} = \{x_1, x_2, \ldots, x_M\}$, the corresponding probability mass function is given by:

$$ P(\mathcal{X} = x) = P(x) = \frac{N(x)}{M}, $$

where $N(x)$ is the number of times the variable $\mathcal{X}$ has value $x$, while $M$ is the total number of values of $\mathcal{X}$. In our case, $\mathcal{X}$ is the $tf$ sequence of a concept $c$ and $P(\mathcal{X} = x)$ is the probability that $tf_c = x$, i.e. the ratio between the number of documents $N(x)$ where a concept appears $x$ times and the total number of documents where $c$ appears, $M$. Given such definition, we denote with $\langle \mathcal{X} \rangle$, $\sigma_{\mathcal{X}}$ and $\langle \ln(\mathcal{X}) \rangle$, respectively: the average, standard deviation and average of the logarithm of $\mathcal{X}$. The algorithm is made by the following steps:

1. Collection of the $tf$:

For each concept $c$, we collect the values of its $tf$ into a list, $l_{tf}$. After that, we compute the standard deviation of the set of values in such list, $\sigma_{l_{tf}}$. If the standard deviation is equal to zero, then it means that either the concept has appeared in only one paper or that it has appeared always the same number of times within the papers. Hence, we discard such concepts since their entropy is zero. For the remaining concepts, we count how many times $tf_c = k \in l_{tf} \forall k$ and store such number into another list named $tf^{\exp}_c$. Thus, the probability that concept $c$ has a given $tf$ is:

$$ p_c^{\exp}(k) = \frac{tf_c^{\exp}(k)}{\sum_{k'} tf_c^{\exp}(k')}. \quad (S22) $$

2. Extraction of fit parameters:

In order to get a power law with a cutoff using the Lagrange multiplier method, we need to impose two constraints: the expected value of $\langle tf \rangle_{th}$ and of $\langle \ln(tf) \rangle_{th}$ have to match the same quantities computed on the data, as discussed in Sec. SIC 1. The analytical form of the maximum entropy distribution becomes:

$$ p(tf_c = k) \equiv p_c^{th}(k) = \frac{1}{Z} \frac{e^{-\lambda k}}{k^n}. \quad (S23) $$

where $Z$ is the normalization constant corresponding to the polylogarithm function $Li_n(e^{-\lambda})$ of order $n$ and argument $e^{-\lambda}$, defined as:

$$ Z \equiv Li_n(e^{-\lambda}) = \sum_{k=1}^{\infty} \frac{e^{-\lambda k}}{k^n}, \quad (S24) $$

$$ e^{-\lambda} = \frac{\langle \ln(tf) \rangle_{th}}{\langle tf \rangle_{th}}. \quad (S25) $$

The computation of $\langle tf \rangle_{th}$ and $\langle \ln(tf) \rangle_{th}$ is made on the first $N_{th}$ concepts in the sorted list $l_{tf}$, which are the ones having the highest value of $tf$. Since we want to approximate the probability distribution function of $\mathcal{X}$ using a power law, we have to impose the constraint $\lambda > 0$, which is the expected case for a power law distribution with a cutoff. Therefore, we impose $\lambda$ to be the value $\lambda_{th}$ obtained in the previous step.

Finally, given a collection of documents $\mathcal{D}$, we compute the expected term-frequency distribution:

$$ tf_c^{th} = \sum_{k=1}^{\infty} tf_c^{\exp}(k) li_n\left(\frac{k}{k_{th}}\right), $$

where $k_{th}$ is the cutoff, typically the $95\%$ value of the distribution $tf_c^{\exp}$.
Theoretical distribution \( p_c^{th}(k) \) depends on two parameters: \( n \) and \( \lambda \). There are two ways to compute their values:

(a) Exploit the fact that the theoretical maximum entropy distribution must reproduce the expectation values \( \langle l_{tf} \rangle \) and \( \langle \ln(l_{tf}) \rangle \). Therefore, we can find \( n \) and \( \lambda \) by solving numerically the following system:

\[
\begin{align*}
\langle l_{tf} \rangle &= \frac{\text{Li}_{n-1}(e^{-\lambda})}{\text{Li}_n(e^{-\lambda})}, \\
\langle \ln(l_{tf}) \rangle &= -\frac{\partial_n \text{Li}_n(e^{-\lambda})}{\text{Li}_n(e^{-\lambda})} = \sum_{k=1}^{\infty} \frac{e^{-\lambda k} k^n \ln(k)}{\text{Li}_n(e^{-\lambda})}.
\end{align*}
\]  

\( (S25) \)

Since the polylogarithm function appears in the above system, we need to use the Python package named mpmath \(^{[62]}\) which implements functions and methods with arbitrary precision float arithmetics. Thus, we define the two equations that have to be solved simultaneously as:

```python
from mpmath import polylog, diff, findroot, fdiv
from math import log as mln
from math import exp as mexp

def eqs(n, z):
    eqA = fdiv(polylog(n-1,z),polylog(n,z)) - avg_tf
    eqB = fdiv(-diff(lambda v: polylog(v,z), n),polylog(n,z)) - avg_ln_tf
    return (eqA, eqB)
```

where \( fdiv \) performs the division in mpmath, while \( \text{diff} \) is used to calculate numerically the derivative of the function \( \text{polylog} \) with respect to \( n \). Then, we use the \textit{findroot} function of mpmath to numerically solve the system of equations with:

```python
sol=findroot(eqs, ci, solver="secant")
```

Typically, the initial values of the parameters are \( ci = (0.5, \text{mexp}(-0.1)) \). The solution of Eq. \( S25 \) is stored in \( sol \), having \( n \) and \( e^{-\lambda} \) as its first and second element. The two parameters, together with the empirical values of \( \langle l_{tf} \rangle \) and \( \langle \ln(l_{tf}) \rangle \), are then passed to the \textit{max$_{\text{ent}}$} function defined below to compute the maximum entropy.

(b) Use the \textit{maximum likelihood estimators} which employs the full data sequence to determine the parameters directly in \( p_c^{th} \), without relying only on two constraints to do so. In this case, following the technique presented in \(^{[63, 64]}\) we use the Python \textit{powerlaw} package to compute the parameters.

3. Computation of Entropies:

Given the parameters \( n \) and \( \lambda \), we can compute the maximum entropy of a concept \( c \) as:

\[
S_{\text{max}} = \ln \left[ \frac{\text{Li}_n(e^{-\lambda})}{\text{Li}_n(e^{-\lambda})} \right] + \lambda \langle tf^{exp} \rangle_c + n \langle \ln(tf^{exp}) \rangle_c.
\]  

\( (S26) \)

which, implemented in Python, reads as follows:

```python
def max_ent(n,z,avg_tf,avg_ln_tf):
    return mln( fp.polylog(n,z) ) - mln(z)*avg_tf + n*avg_ln_tf
```

The empirical entropy, \( S_c \), is instead computed using Shannon formula (Eq. \( 1 \)) and distribution \( p_c^{exp}(k) \).

2. Density of \( tf \)

As commented previously (see Sec. \textit{SI C 2}), the maximum entropy distribution associated to the case of a rescaled term-frequency sequence, \( rtf \), is a lognormal, defined as:

\[
p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi} \sigma x} \exp \left[ -\left( \frac{\ln(x) - \mu}{2\sigma^2} \right)^2 \right] \quad \text{with} \; x > 0.
\]  

\( (S27) \)
Given a sequence of $M$ continuous values $\mathcal{X} = \{x_1, x_2, \ldots, x_M\}$, we define the probability to observe a value between $x$ and $x + \Delta x$ as $P(x, x + \Delta x)$. To compute such quantity, we have to consider the probability density function $p(x)$ and integrate it across the interval, such that:

$$P(x, x + \Delta x) = \int_x^{x+\Delta x} p(y) \, dy.$$  \hfill (S28)

Under this assumption, the algorithm is made by the following steps:

1. **Collection of $rtf$:**
   For each concept $c$, collect its $rtf$ values into a list, $l_{rtf}$. In analogy with the case of discrete $tf$, we discard those concepts having $\max(rtf) - \min(rtf) \leq 0.005$. Then, we create a binning $\{\Delta k\}$ of the interval $[\min(rtf), \max(rtf)]$ and compute the empirical probability, $P$, that the $rtf$ assumes a value between $k$ and $k + \Delta k$, using Eq. S28.

2. **Extraction of fit parameters:**
   Since the form of the lognormal distribution, Eq. S27, the parameters $\mu$ and $\sigma$ that determine it are directly calculated from the empirical $rtf$ list, $l_{rtf}$, as $\mu \equiv \langle \ln(l_{rtf}) \rangle$ and $\sigma \equiv \sigma(\ln(l_{rtf}))$, where the last term denotes the standard deviation of the logarithm of the term-frequency density $l_{rtf}$.

3. **Computation of the residual entropy:**
   After obtaining parameters $\mu$ and $\sigma$, we compute the residual entropy, $S_d$, using a discrete version of the Kullback-Leibler divergence given by:

$$S_d = \sum P(k, k + \Delta k) \ln \frac{P(k, k + \Delta k)}{Q(k, k + \Delta k)} \Delta k,$$ \hfill (S29)

where the sum is performed over the set of intervals used for the binning $\{\Delta k\}$. It is worth stressing that such binning is the same for both $P$ and $Q$. Such operation is achieved by the following code:

```python
def discrete_KL(data_distro, th_distro, bin_widths):
    return np.sum(data_distro * np.log(np.true_divide(data_distro, th_distro)) * bin_widths)

num_bins_fixed_kl = 15
binning = np.logspace(np.log10(np.min(rescaled_tfs)*0.999), np.log10(np.max(rescaled_tfs)*1.001), num_bins_fixed_kl + 1)

vs_r_tfs, bs_r_tfs = np.histogram(r_tfs, bins = binning, density=True)
centers_bins = (binning[1:] + binning[:-1]) / 2.
bin_ranges = binning[1:] - binning[:-1]

# Removal of bins with no data for the experimental distro
indx_nnz_vs_r_tfs = np.nonzero(vs_r_tfs)
vs_r_tfs_nnz = vs_r_tfs[indx_nnz_vs_r_tfs]
centers_bins_nnz = centers_bins[indx_nnz_vs_r_tfs]
bin_ranges_nnz = bin_ranges[indx_nnz_vs_r_tfs]

# Only calculated for the middle point of the bins for nonzero integral
# values of the data histogram
th_pdf = lognorm.pdf( centers_bins_nnz, loc=0, s=sigma, scale=scale )
dKL = discrete_KL( vs_r_tfs_nnz, th_pdf, bin_ranges_nnz )
```

data_distro and th_distro contain the values of the probability distribution functions evaluated at the center of the intervals $\{\Delta k\}$ for the observed sequence $l_{rtf}$ and the theoretically expected one.
3. Generation of similarity networks

After computing the $S_{max}$, for each concept $c$, we compute its residual entropy, $S_d$, and store its value on a list $l_{sdiff}$. We then compute the percentiles of $l_{sdiff}$ using the numpy function named `percentile`. We compute the percentiles $\tilde{P}$ from $\tilde{P}_{min}$ to $\tilde{P}_{max}$ using a step of $\tilde{P}_{step}$, and use those values to create separate files containing the lists of concepts having $S_d \geq \tilde{P}$. Finally, we build the similarity network between documents using exclusively those concepts contained in the file corresponding to the $i$-th percentile $\tilde{P}$.

[56] Kullback S, Leibler RA (1951) On information and sufficiency. The annals of mathematical statistics 22(1):79-86.
[57] D. Hric, R. K. Darst, and S. Fortunato (2014), Community detection in networks: Structural communities versus ground truth, Phys Rev E, 90:062805.
[58] arXiv e-print service. Available at [https://arxiv.org/](https://arxiv.org/).
[59] The Twitter Developer Documentation, API overview. Available at [https://dev.twitter.com/overview/api](https://dev.twitter.com/overview/api).
[60] Python Software Foundation. Python Language Reference (version 2.7). Available at [http://www.python.org](http://www.python.org).
[61] Jones E, Oliphant T, Peterson P et al. and Plenz D (2001) SciPy: Open source scientific tools for Python. Available at [https://www.scipy.org](https://www.scipy.org).
[62] Fredrik Johansson et al. (2013) mpmath: a Python library for arbitrary-precision floating-point arithmetic (version 0.18). Available at [http://mpmath.org/](http://mpmath.org/).
[63] Clauset A, Shalizi CR and Newman MEJ (2009) Power-law distributions in empirical data. SIAM Review 51(4):661-703.
[64] Alstott J, Bullmore E and Plenz D (2014) Powerlaw: a Python package for analysis of heavy-tailed distributions. PLoS ONE 9(1):e85777. Available at [https://pypi.python.org/pypi/powerlaw](https://pypi.python.org/pypi/powerlaw).