Unleashing the power of semantic text analysis: a complex systems approach

THÈSE N^{O} 8473 (2018)

PRÉSENTÉE LE 16 MARS 2018

À LA FACULTÉ DES SCIENCES DE BASE
LABORATOIRE DE BIOPHYSIQUE STATISTIQUE
PROGRAMME DOCTORAL EN PHYSIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Andrea MARTINI

acceptée sur proposition du jury:

Prof. H. M. Rønnow, président du jury
Prof. P. De Los Rios, directeur de thèse
Prof. A. Flammini, rapporteur
Prof. J. Gómez Gardeñes, rapporteur
Prof. R. West, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2018

When there is so much to be known,
when there are so many fields of knowledge
in which the same words are used with different meanings,
when every one knows a little about a great many things,
it becomes increasingly difficult for anyone to know
whether he knows what he is talking about or not.
And when we do not know, or when we do not know enough,
we tend always to substitute emotions for thoughts.

— Thomas Stearns Eliot

A Nuccia e Silvio,

ad Ahmadreza Djalali.

Acknowledgements

I apologize for not using English here, but I prefer to write what I have to say directly in the language that I used to communicate with the recipients.

I primi ringraziamenti, chiaramente, sono per Paolo, il mio supervisore. Lo ringrazio per avermi dato l'opportunità di intraprendere il mio dottorato: dal punto di vista scientifico, il suo modo di pensare e le domande che si pone mi hanno fatto capire in profondità quale sia un buon metodo per attaccare criticamente i problemi scientifici che ci si trova ad affrontare. Dal punto di vista umano, i tanti momenti di convivialità, le chiaccherate e le tante risate che ci siam fatti non hanno potuto che portarmi a considerarlo più come un amico che come un capo.

Ringrazio inoltre Alessio, che ha condiviso con me il lavoro per la quasi totalità di questo percorso. Per la sua completa disponibilità a rompergli le scatole; per tutti gli infiniti consigli utili su come affrontare i problemi scientifici da fisico, tenendo bene in mente qual'è il contributo che si può dare in quest'ottica; per aver dedicato molto tempo nel supportare il mio lavoro, insegnandomi l'arte delle presentazioni; infine, per la smisurata pazienza che ha avuto in innumerevoli occasioni in cui, nonostante tutto, mi ha sempre trattato con rispetto e gentilezza. Per tutto questo, posso dire che è stato il mio co-supervisore, anche se non a livello ufficiale. Inoltre, lo ringrazio anche dal punto di vista umano: inevitabilmente, mi sono molto divertito assieme (ormai i programmi di satira che furono ce li ricordiamo noi due, a vicenda) e si è venuto a creare un bel sentimento di amicizia.

Grazie i compagni (di ufficio), Alberto e Duccio, per aver condiviso lo spazio-tempo lavorativo assieme, oltre a tante discussioni sui temi più disparati.

A Davide e Peppe, con cui condivido il migliore coinquilinaggio che abbia mai avuto, nonostante la mia presenza effemira durante la settimana.

Gracias a Julia, por su cariño y su buen rollo, por las muchas copas y por sus cenas deliciosas, y por "aguantar" ese guiris.

Gracias a Raquel, porqué no parece que han pasado tantos años sin vernos y siempre tenemos el mismo cariño.

Grazie alla combriccola italiana a Losanna, per le tante sontuose mangiate che abbiam fatto

assieme e per quelle a venire.

Grazie a mamma, papà e Gabri, per essermi venuti a trovare in terra straniera ed per avermi sempre appoggiato nelle mie scelte, credendo appieno nelle mie ricchezze e dandomi preziosi consigli.

Grazie agli amici della valle tutt*, che anche se ci scriviamo poco e ci vediamo ancora meno, rimanete un punto fisso per il supporto che mi date e per il bene che vi voglio, il tutto sempre e allegramente con la pancia piena ed il bicchiere anche.

Lausanne, 17 December 2017 A. M.

Abstract

In the present information era, a huge amount of machine-readable data is available regarding scientific publications. Such unprecedented wealth of data offers the opportunity to investigate science itself as a complex interacting system by means of quantitative approaches. These kind of studies have the potential to provide new insights on the large-scale organization of science and the driving mechanisms underlying its evolution. A particularly important aspect of these data is the semantic information present within publications as it grants access to the concepts used by scientists to describe their findings. Nevertheless, the presence of the so-called buzzwords, *i.e.* terms that are not specific and are used indistinctly in many contexts, hinders the emerging of the thematic organization of scientific articles.

In this Thesis, I resume my original contribution to the problem of leveraging the semantic information contained in a corpus of documents. Specifically, I have developed an information-theoretic measure, based on the maximum entropy principle, to quantify the information content of scientific concepts. This measure provides an objective and powerful way to identify generic concepts acting as buzzwords, which increase the noise present in the semantic similarity between articles. I prove that the removal of generic concepts is beneficial in terms of the sparsity of the similarity network, thus allowing the detection of communities of articles that are related to more specific themes. The same effect is observed when describing the corpus of articles in terms of topics, namely clusters of concepts that compose the papers as a mixture. Moreover, I applied the method to a collection of web documents obtaining a similar effect despite their differences with scientific articles. Regarding the scientific knowledge, another important aspect I examine is the temporal evolution of the concept generality, as it may potentially describe typical patterns in the evolution of concepts that can highlight the way in which they are consumed over time.

Keywords: Complex systems | science of science | semantic networks | community detection | topic modeling | maximum entropy principle | applied statistical physics

Sommario

Nell'era dell'informazione nella quale viviamo, una grande quantità di dati riguardanti le pubblicazioni scientifiche risulta disponibile in un formato che può essere facilmente trattato da un computer. Tale mole di dati, la cui grandezza è senza precedenti, offre l'opportunità di investigare la scienza stessa nell'ottica dei sistemi complessi tramite approcci quantitativi. Questa tipologia di studi ha il potenziale per fornire nuove conoscenze sull'organizzazione a grande scala della scienza, e sui meccanismi che fanno da motore alla sua evoluzione. Un aspetto particolarmente importante di questi dati è l'informazione semantica presente all'interno delle pubblicazioni, che garantisce l'accesso ai concetti utilizzati dagli scienziati per descrivere i risultati della loro ricerca. Tuttavia, la presenza di concetti popolari, cioè termini che non sono specifici e vengono usati indistintamente in molti contesti, impedisce l'emersione dell'organizzazione tematica degli articoli scientifici.

Nella presente Tesi riassumo il mio apporto originale al problema dell'utilizzo dell'informazione semantica contenuta negli articoli scientifici. In special modo, ho sviluppato una misura di teoria dell'informazione fondata sul principio di massima entropia per quantificare il contenuto informativo dei concetti scientifici. Tale misura rappresenta un metodo oggettivo e potente per identificare quei concetti generici che agiscono come termini dal significato vago, accrescendo il "rumore" presente nelle reti di similarità semantica tra articoli. Inoltre, dimostro che la rimozione dei concetti generici porta dei benefici in termini di riduzione della densità della rete di similaritá, permettendo di identificare gruppi di articoli che trattano specifici temi. Lo stesso effetto si osserva se gli articoli vengono descritti come composti da un misto di diversi temi costituiti da gruppi di parole. Il metodo è stato anche applicato ad un gruppo di documenti web ottenendo un effetto simile nonostante questi siano chiaramente differenti rispetto agli articoli scientifici per struttura e contenuti. Per quel che riguarda la conoscenza scientifica, un'altro aspetto importante che ho esaminato è stata l'evoluzione nel tempo della generalitá dei concetti, dato che può essere utilizzata per descrivere degli andamenti tipici nell'evoluzione dei concetti che possono evidenziare il modo in cui essi vengono utilizzati al variare del tempo.

Parole chiave: sistemi complessi | scienza della scienza | reti semantiche |
identificazione dell'organizzazione in comunità | modelizzazione dei temi |
fisica statistica applicata

Contents

List of Figures

List of Tables

Introduction

The present-day interest in characterizing and understanding the driving mechanisms of the scientific production is placed within the general curiosity in the comprehension of human activities that range from the use of social media to increase people awareness [1] to the description of recurrent mobility patterns [2, 3]. Scientists are increasingly attracted by the narrative of science, especially if approached from a global viewpoint in order to discover large-scale trends that guide such collective effort [4]. Despite a widespread enthusiasm in such studies surfaced only recently [5], a longstanding tradition in the inquiry of scientific knowledge and how it evolves constitutes a central part of the philosophy and the history of science [6, 7]. These fields of research are indeed essential to provide a systemic overview of scientific paradigms and milestones from a qualitative point of view.

On the other hand, quantitative analysis of science dates back to 1930s with the publication of the book of John D. Bernal entitled *The Social Function of Science* [8]. However, it was only after World War II that the interest in the field began to flourish [9]. The first commercial release of citation indexes in 1950s [10] immediately triggered the enthusiasm toward the opportunities that the availability of such information opened up: the seminal paper of Eugene Garfield, published in 1955, was a precursor of the modern citation analysis [11]. Garfield itself founded the renowned Institute of Scientific Information (ISI) and prompted the development of the Science Citation Index, a database collecting information about scholar manuscripts from several disciplines [12]. A decade after Garfield's paper, de Solla Price published two books, *Science Since Babylon* and *Little Science, Big Science*, between 1961 and 1963, and right after the milestone paper in scientometrics, entitled *Networks of scientific papers* [13–15].

Although, in their prime, this kind of studies were the focus of the research activity of a small community of scholars, they involve nowadays a much bigger community of scientists coming from a broader range of disciplines. The gathering of all these studies/scientists gave rise to a new domain called *science of science*. This term broadly indicates all the works that investigate some *facet of science*[1] in the framework of the scientific method without restricting to quantitative studies (as opposed to qualitative) or to natural sciences only [16]. Accordingly, various scientific

[1] The definition of 'science' in this domain is customarily intended in its full breath as including any "... ordered and reliable knowledge – so that a philologist or a critical historian can truly be called scientific ... " (taken from [16], p. 390) as quoted in [5].

disciplines contribute to the study of science of science, ranging from the sociology of science to its mathematical modeling, to just mention a few. This wealth of cultures, each one bringing in its own approaches and methods, implies that science of science did not become a self-consistent branch of knowledge like economics or sociology, but the studies are developing somehow independently as small niches within various fields [5].

In the present era of *data deluge*, the science of science has been positively affected by the advances in information technology. In particular, the digitalization of scientific communication is providing a wealth of data regarding scientific publications [17–21]. Aside from that, the number of publications keeps growing as a consequence of the rise in the number of scientists which is primarily induced by the higher educational level of population and the recognized benefits of science for society [22–29]. Furthermore, modern science itself is experiencing a pressure toward specialization that enhances the sophistication of instruments (think about the Large Hadron Collider at CERN), experimental techniques, analytical methods and theories [30–33]. On the one hand, the combination of these continuous changes with an ever-increasing rate of publications makes dealing with such overwhelming information not trivial. On the other hand, the massive amount of publication data that can be processed offers the opportunity to confirm conjectures and to distillate innovative insights supported by quantitative analyses at unprecedented scales, both time- and domain-wise. The interpretation of empirical patterns ultimately allows to answer fundamental questions about science like *which are the driving mechanisms that cause the emergence of research fronts and the shifts of scientific paradigms* [34–36]. Nevertheless, it becomes more problematic to extract reliable information as publication data are usually noisy or incomplete. Both the low quality and the unstructured heterogeneity of these data pose practical challenges that undermine their potential applications, contributing to dilute the efforts toward the possibility to make general claims about science.

The information available in a scientific article ideally contains the journal of publication, the list of authors and their affiliation, the semantic component (title, abstract and text), and references to cited articles [37]. Moreover, some journal provide the classification of a paper in different subjects and important keywords usually selected by the authors. Among the many uses of the various information about publications, relevant examples include the investigation of the patterns of co-authorships [38–42], the influence of the geographical position of the institutions on the scientific collaborations and individual careers [43–49], the evolution of the impact of papers [50–55], and the characterization of the scientific activity of scholars [56–59].

One approach to gain knowledge from such information is to construct a map of science at the article level, analyzing the relationships between them at a coarse level. The reasonable assumption behind this approach is that it should exist an underlying organization of science in different disciplines and fields [60, 61], where the role and effects of such organization should manifest when analyzing some of the entities that articles possess. The most exploited facet is the similarity of the reference lists in the citation space, which leverages the pattern of citations among articles [62–64]. All the above studies, however, base their findings on the bibliographic information associated with articles *i.e.* authors, affiliations, and citations. As a matter of fact,

most of the information contained in the articles is actually overlooked. Clearly, the vast majority of an article is constituted by the text that contains the explanation of the research and the analysis of its outcomes. Ignoring this kind of information is not really efficient as we are not considering the message that the authors want to convey and the topic of research that they address, which represent the reason why the paper was written, *i.e.* to disseminate its content. Considering the semantic content of articles is a different way to map the scientific knowledge [65–67]. In this way, we should be able to identify topically related articles that share concepts, methods or ideas, potentially uncovering similarities in their content at a broader semantic level that goes beyond citations [63, 68–70]. Moreover, a careful analysis of the topic structure is also useful to characterize the specialization of science. However, understanding the thematic organization of a corpus has been a longstanding effort also in information retrieval, mainly related to the search of relevant documents in the vast literature of a domain [71, 72]. Contributions from text analysis and information retrieval scholars allowed to define the (current) standard approach to extract the subjects within a corpus of documents going under the name of *topic modeling*. This approach aims at describing documents as composed by topics, *i.e.* groups of semantically related words that co-occur more frequently together [73]. Words in single documents are then modeled as if they are extracted from the topics that compose the documents. Indeed, the hypothesis behind topic modeling is that a latent thematic division is present in the corpus and, therefore, the model adopted should be able to highlight it automatically.

Generally speaking, the studies on science of science leverage various types of information associated to scientific publications. The different perspectives that can be considered when analyzing these data ideally place these researches in the domain of *complexity science* [74, 75]. Indeed, a system that is composed by several kind of constituents related to each other in very different ways is well suited to be described as a *complex system*. Given the multitude of constituents and interactions, a compact but complete description of such systems can be provided by modeling the interactions between constituents as a network. This formalism allows to looking through the intrinsic complexity of these systems, characterizing the emergence of interesting phenomena from the collective interaction of the constituents. The analysis of the collective effects in interacting systems is the subject of *statistical physics*, a fundamental branch of Physics that investigate systems with many constituents where macroscopic phenomena, like the self-organization of articles in topics, can develop from microscopic properties, *e.g.* the way concepts are adopted in single articles. Therefore, a physicist examining complex systems is certainly intrigued by their faceted nature and will likely be excited by the possibility of characterizing the relationships between their constituents from the micro- to the macro-scale.

The objective of the present Thesis is to identify the patterns of organization emerging in scientific knowledge using methodologies grounded on statistical physics. To this aim, we focused on the analysis of the semantic content of scientific articles. More specifically, we design a method intimately connected to the notion of entropy in information theory in order to detect relevant concepts within articles. Linking concepts with such feature allows to discern their role in shaping the large-scale organization of scientific knowledge when examined under the lens of the topic composition. This perspective is indeed important to consider since the growing number of

scientific articles poses serious challenges to the scalability of traditional classification schemes that rely on human supervision. The design of effective methods that characterize automatically the composition of a corpus of scientific articles is, therefore, a vivid exigence. Apart from the scientific interest in exploring the semantic organization of knowledge *per se*, the practical implications of such methods are even more pressing as navigating the existing literature to find a paper of interest is is becoming increasingly difficult due to the fast growth of available publications [76–78]. Within the same perspective, we study the evolution in time of the "role" of a scientific concept to identify trends and associate them to different roles. For example, hot research topics may popularize the adoption of concepts that were previously used only by a restricted community of scholars or, on the contrary, it may happens that other concepts become obsolete after being commonly embraced.

The present Thesis is organized as follows: in chapter 1 we gently introduce from scratch the conceptual definitions that will be used throughout the rest of the Thesis. Apart from the setup of a common terminology, we also describe various methodologies, techniques and measures, providing adequate reasonings and motivations that lead to their adoption. In particular, section 1.1 is devoted to the fundamentals of network theory, and section 1.2 to the notion of information entropy. In chapter 2 we apply concepts from network theory to study the semantic similarity between documents, conceiving a method to establish the relevance of scientific concepts based on entropy. Such method provides a well-grounded criterion to remove blurring noise in the paper similarities, a thoughtful operation that proves to be beneficial – among other things – for ameliorating the organization of the network in groups of thematically related articles. In chapter 3 we investigate the evolution in time of concept relevance, addressing the possibility to detect common trends in their history. Ideally, collecting concepts with related trajectories would allow to analyze their common fate, highlighting similarities and differences between the roles that they take on concerning the knowledge transformations. Finally, in the conclusions we summarize the main results that have been obtained, pointing out the original contribution of our research in the framework of existing knowledge about science of science.

1 Methods

1.1 Introduction to complex networks

Networks, also called **graphs** [79–84], provide a general yet simple formalism to describe the interactions among constituents in real-world systems, ranging from social structures [85–90] to technological infrastructures [91–97], biological entities [98–103] and physical systems [104–106]. Despite the heterogeneous origin of such systems, they consist of elementary components interacting together in a non-trivial fashion: no matter the peculiar features of the single constituents or the nature of their interactions, they all give rise to the emergence of collective phenomena that are surprisingly similar and cannot be explained as the pure sum of the actions of individual units [107]. Such aspect is the distinctive hallmark of **complex systems** where the interactions among the units drive the spontaneous emergence of an organization. Therefore, systems whose interaction patterns are very much alike and are characterized by non-trivial topologies can be described as **complex networks**.

In the network framework, the basic units that compose a system are denoted as **nodes** (or **vertices**) connected to each other by **links** (or **edges**) that symbolize the interactions among them, as depicted in Figure 1.1. This abstract description of complex networks can be formalized in mathematical terms by means of **graph theory** [108, 109]. In particular, we resort to it focusing on a specific kind of graphs known as **bipartite networks**.

1.1.1 Bipartite networks

Imagine that we want to characterize a system constituted by two distinct types of entities where relationships exist only between entities of different type [110–115]. An example of such network is provided by the affiliation network of actors and movies, where actors are solely linked to movies they participated in [116–118]. These kinds of systems are best represented by a **weighted bipartite graph** $\mathcal{B} = \left(\mathcal{V}_{\mathcal{P}}, \mathcal{V}_{\mathcal{Q}}, \mathcal{Z}\right)$, a mathematical object composed by two independent sets $\mathcal{V}_{\mathcal{P}}$ and $\mathcal{V}_{\mathcal{Q}}$, each containing entities of the same type as elements, and a third set $\mathcal{Z} = \left\{z(v_p, v_q) \in \mathbb{R}_+ | v_p \in \mathcal{V}_{\mathcal{P}}, v_q \in \mathcal{V}_{\mathcal{Q}}\right\}$, which elements are positive numbers that indicate the intensity

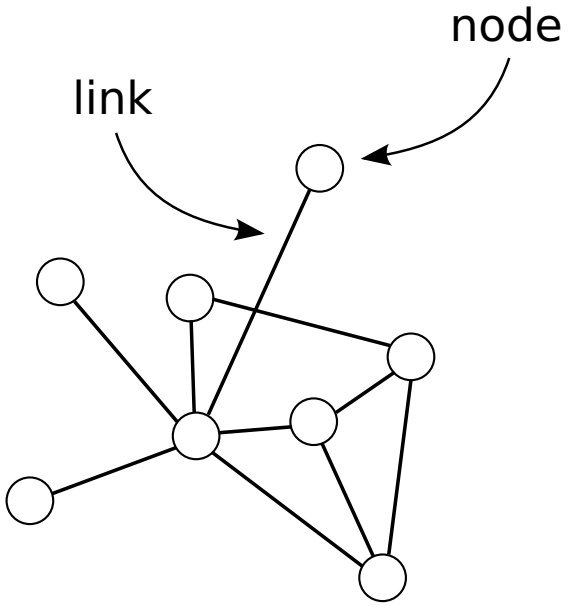

Figure 1.1 – Sketch of a network where nodes are represented by circles and links are drawn as lines between pairs of nodes. The network is composed by 10 links and 8 nodes.

of the interaction between elements taken from $\mathcal{V}_\mathcal{P}$ and $\mathcal{V}_\mathcal{Q}$ respectively. Since the two sets $\mathcal{V}_\mathcal{P}$ and $\mathcal{V}_\mathcal{Q}$ are independent, their intersection is the empty set, $\mathcal{V}_\mathcal{P} \cap \mathcal{V}_\mathcal{Q} = \emptyset$. Moreover, entities that are not interacting do not have a representative element in the set $\mathcal{Z}$. The cardinality, *i.e.* the number of elements, of the three sets is thus $\left|\mathcal{V}_\mathcal{P}\right| = P$, $\left|\mathcal{V}_\mathcal{Q}\right| = Q$ and $\left|\mathcal{Z}\right| = E$.

In the network language, the elements of $\mathcal{V}_\mathcal{P}$ and $\mathcal{V}_\mathcal{Q}$ are the **nodes** of distinct type of the bipartite network while the elements of $\mathcal{Z}$ are the **link weights**. Any two nodes $v_p \in \mathcal{V}_\mathcal{P}$ and $v_q \in \mathcal{V}_\mathcal{Q}$ connected by a link are denoted as **adjacent** or **neighbors** and the set of neighbors of a given node v_p is called the **neighborhood** of v_p. A compact way to describe the bipartite graph $\mathcal{B}$ is through its **weighted biadjacency matrix** [110], which is the $P \times Q$ matrix $\boldsymbol{A}_\mathcal{B}$ where the entry in row p and column q, a_{pq}, is equal to the link weight $z(v_p, v_q)$ that connects two adjacent nodes v_p and v_q of the network and zero otherwise. An illustrative instance of a weighted bipartite graph $\mathcal{B}$ is displayed in Figure 1.2 (a) together with its biadjacency matrix $\boldsymbol{A}_\mathcal{B}$ in Figure 1.2 (b). Although the bipartite networks formalism is the natural framework to describe systems where interactions occur only between constituents of two different types, it is often easier to analyze the property of a system in terms of relationships among constituents of a single type. The procedure that we adopt to compute such relationships is called one-mode or unipartite projection of a bipartite network [119]. Such projection onto a one-mode network implies that there is always a loss of information, though it can be mitigated by an appropriate weighting of the links in the resulting unipartite network [120, 121]. In the following part, we provide the mathematical details of the projection scheme [122].

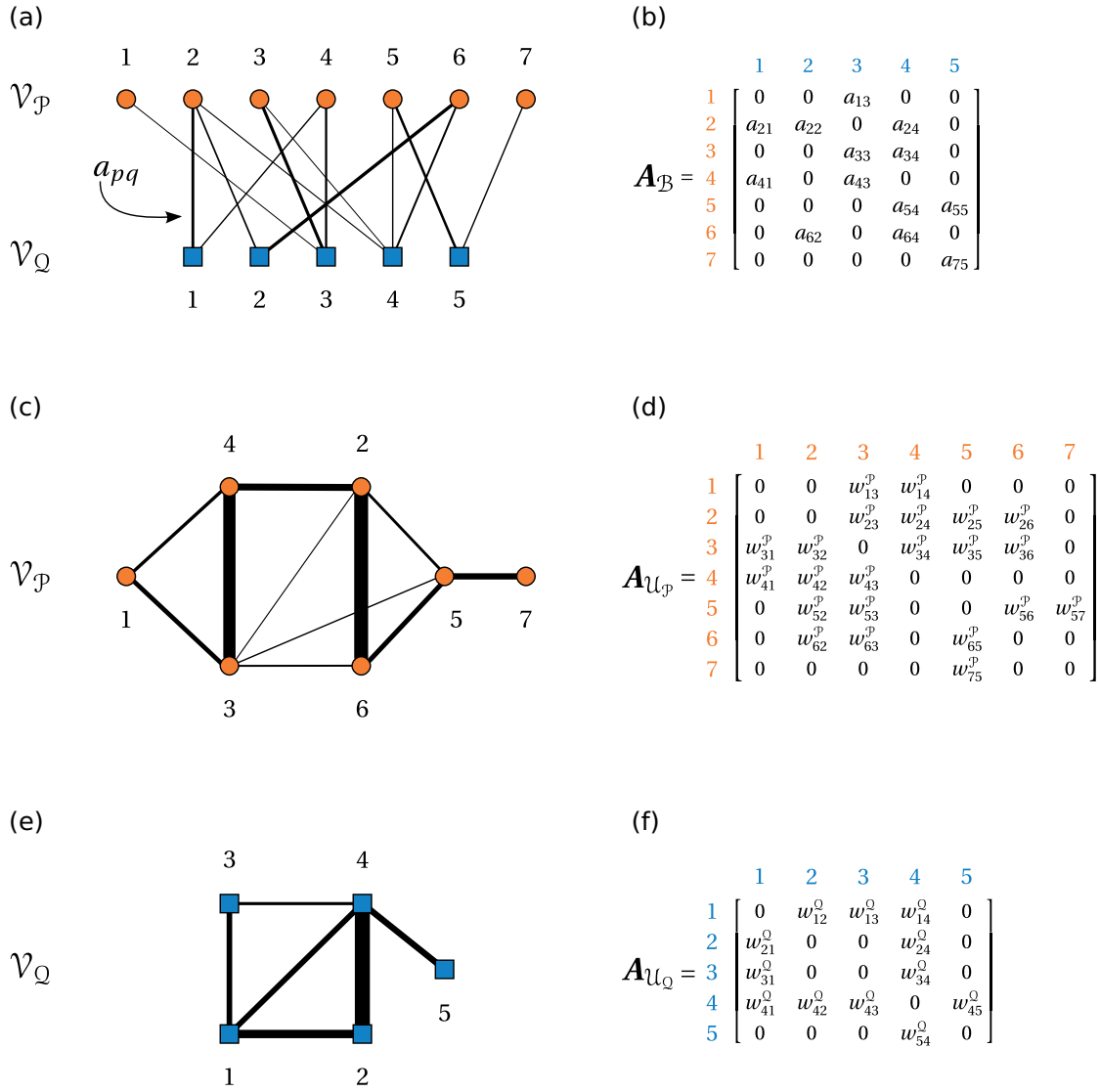

Figure 1.2 – Prototype of a bipartite network $\mathcal{B}$ consisting of $P = 7$ nodes from set $\mathcal{V}_\mathcal{P}$ and $Q = 5$ nodes from set $\mathcal{V}_\mathcal{Q}$ with a total number of links $E = 13$. A schematic layout is illustrated in (a) where nodes in $\mathcal{V}_\mathcal{P}$ and $\mathcal{V}_\mathcal{Q}$ are represented by orange circles and blue squares, respectively, while links are the lines that run between them. The weight of the link between nodes v_p and v_q is indicated as a_{pq} and graphically displayed by the thickness of the line. The biadjacency matrix $\boldsymbol{A}_\mathcal{B}$ of the graph is presented in (b): rows and columns indices are colored according to the set of nodes that they represent in (a), *i.e.* $\mathcal{V}_\mathcal{P}$ and $\mathcal{V}_\mathcal{Q}$ respectively. The unipartite projection onto the nodes in $\mathcal{V}_\mathcal{P}$ is shown in (c) along with its weighted adjacency matrix $\boldsymbol{A}_{\mathcal{U}_\mathcal{P}}$ outlined in (d), where the link weight between nodes i and j, $w^\mathcal{P}_{ij}$, is calculated from $\boldsymbol{A}_\mathcal{B}$ according to Equation 1.1. Similarly, the unipartite projection onto the nodes in $\mathcal{V}_\mathcal{Q}$ is shown in (e) along with its weighted adjacency matrix $\boldsymbol{A}_{\mathcal{U}_\mathcal{Q}}$ outlined in (f), where the link weight between nodes i and j, $w^\mathcal{Q}_{ij}$, is calculated from $\boldsymbol{A}_\mathcal{B}$ according to Equation 1.2.

1.1.2 Unipartite projection of bipartite networks

The **unipartite projection of a bipartite network** $\mathcal{B} = (\mathcal{V}_\mathcal{P}, \mathcal{V}_\mathcal{Q}, \mathcal{Z})$ is defined as a **weighted graph** $\mathcal{U}_\mathcal{N} = (\mathcal{V}_\mathcal{N}, \mathcal{W}_\mathcal{N})$ where elements of the set $\mathcal{V}_\mathcal{N}$ are nodes taken exclusively from one of the two sets of nodes, *i.e.* either $\mathcal{V}_\mathcal{N} = \mathcal{V}_\mathcal{P}$ or $\mathcal{V}_\mathcal{N} = \mathcal{V}_\mathcal{Q}$. The **size** of the unipartite graph is the number of nodes in the network, namely the cardinality of the set of nodes $|\mathcal{V}_\mathcal{N}| = N$. The elements of the set $\mathcal{W}_\mathcal{N}$ are the weights of the links among nodes in $\mathcal{V}_\mathcal{N}$, *i.e.* $\mathcal{W}_\mathcal{N} = \left\{ w_{ij}^\mathcal{N} \in \mathbb{R}_+ | v_i, v_j \in \mathcal{V}_\mathcal{N} \right\}$, where a link between two nodes $v_i, v_j \in \mathcal{V}_\mathcal{N}$ exists only if they share at least one adjacent node in the bipartite network $\mathcal{B}$. For example, consider the bipartite network $\mathcal{B}$ in Figure 1.2 (a): if we take the projection onto the one-mode network composed by nodes in $\mathcal{V}_\mathcal{P}$, nodes 4 and 5 do not have any adjacent node in common in $\mathcal{B}$. As a consequence, these nodes are not adjacent in the projected network $\mathcal{U}_\mathcal{P}$. The link weight $w_{ij}^\mathcal{N}$ between nodes v_i and v_j is thus defined as the sum of the product between link weights in $\mathcal{Z}$ that join common adjacent nodes of v_i and v_j in the bipartite network $\mathcal{B}$. Although other policies to compute the link weights are possible [119], here we adopt this simple definition. The projected network $\mathcal{U}_\mathcal{N}$ can be fully described in a concise form by means of the so-called **weighted adjacency matrix** $\boldsymbol{A}_{\mathcal{U}_\mathcal{N}}$ [123]. Such matrix is akin to the biadjacency matrix $\boldsymbol{A}_\mathcal{B}$ since it contains the intensity of the interactions among nodes in $\mathcal{V}_\mathcal{N}$. As a consequence, it is an $N \times N$ matrix. The entry in row i and column j denotes the link weight $w_{ij}^\mathcal{N}$ from node v_i to node v_j. In general, such interactions are not symmetric, $w_{ji}^\mathcal{N} \neq w_{ij}^\mathcal{N}$. In this case, since the order of the nodes is important, we can associate a direction to the link that goes from a source node to a target node [124]. Conversely, if there is no reason to assume that the interactions are directed, the weighted adjacency matrix is symmetric, *i.e.* the entry in row i and column j is the same as in row j and column i, $w_{ij}^\mathcal{N} = w_{ji}^\mathcal{N}$. Moreover, self-loops that connect nodes to themselves are not usually allowed, therefore the diagonal entries $w_{ii}^\mathcal{N}$ are set to zero.

If we consider the unipartite projection $\mathcal{U}_\mathcal{P}$ that consist of nodes in $\mathcal{V}_\mathcal{P}$, the entry $w_{ij}^\mathcal{P}$ is defined in terms of the entries of the biadjacency matrix $\boldsymbol{A}_\mathcal{B}$ as follows:

$$w_{ij}^\mathcal{P} = a_{i1}a_{j1} + a_{i2}a_{j2} + \cdots + a_{iQ}a_{jQ} = \sum_{q=1}^{Q} a_{iq}a_{jq} = \boldsymbol{a}_{i:} \cdot \boldsymbol{a}_{j:} \qquad \left(i, j = 1, \ldots, P\,, \quad i \neq j\right), \tag{1.1}$$

where $\boldsymbol{a}_{i:}$ and $\boldsymbol{a}_{j:}$ are shorthands for the vectors that represent rows i and j of $\boldsymbol{A}_\mathcal{B}$, respectively, while the symbol $\cdot$ denotes the dot product between vectors. The relation in Equation 1.1 is nothing else than the matrix product between the biadjacency matrix $\boldsymbol{A}_\mathcal{B}$ and its transpose $\boldsymbol{A}_\mathcal{B}{}'$, $w_{ij}^\mathcal{P} = \left(\boldsymbol{A}_\mathcal{B}\boldsymbol{A}_\mathcal{B}{}'\right)_{ij} = \sum_{q=1}^{Q} a_{iq}a'_{qj} = \sum_{q=1}^{Q} a_{iq}a_{jq}$, where the entry in row q and column j of the latter is equal to the entry in row j and column q of the former, $a'_{qj} = a_{jq}$. As an example, the one-mode projected network $\mathcal{U}_\mathcal{P}$ obtained from the bipartite network $\mathcal{B}$ in Figure 1.2 (a) is displayed in Figure 1.2 (c), along with its weighted adjacency matrix $\boldsymbol{A}_{\mathcal{U}_\mathcal{P}}$ in Figure 1.2 (d).

In the same manner, we can examine the unipartite projection $\mathcal{U}_\mathcal{Q}$ composed by nodes in $\mathcal{V}_\mathcal{Q}$,

where the entry $w^{\mathcal{Q}}_{ij}$ of its weighted adjacency matrix $\boldsymbol{A}_{\mathcal{U}_\mathcal{Q}}$ is

$$w^{\mathcal{Q}}_{ij} = \sum_{p=1}^{P} a_{pi}\, a_{pj} = \boldsymbol{a}_{:i} \cdot \boldsymbol{a}_{:j} \qquad \left(i, j = 1, \ldots, Q\,, \quad i \neq j\right), \tag{1.2}$$

where $\boldsymbol{a}_{:i}$ and $\boldsymbol{a}_{:j}$ are shorthands for the vectors that represent columns i and j of $\boldsymbol{A}_{\mathcal{B}}$, respectively. Again, the weight $w^{\mathcal{Q}}_{ij}$ can be recast as the matrix product between the transpose of the biadjacency matrix $\boldsymbol{A}_{\mathcal{B}}{}'$ and the biadjacency matrix $\boldsymbol{A}_{\mathcal{B}}$ itself, $w^{\mathcal{Q}}_{ij} = \left(\boldsymbol{A}_{\mathcal{B}}{}'\boldsymbol{A}_{\mathcal{B}}\right)_{ij} = \sum_{p=1}^{P} a'_{ip} a_{pj} = \sum_{p=1}^{P} a_{pi} a_{pj}$ [89, 125]. In this case, the one-mode projected network $\mathcal{U}_\mathcal{Q}$ and the corresponding adjacency matrix $\boldsymbol{A}_{\mathcal{U}_\mathcal{Q}}$ are illustrated in Figure 1.2 (e) and (f), respectively.

After having introduced how to project a weighted bipartite network onto a unipartite one, we take a step further in order to analyze its interaction pattern. Indeed, the characterization of various facets of the interactions is of great importance to describe the large-scale behavior of the systems encoded as networks.

1.1.3 Topological indicators of unipartite networks

The weighted adjacency matrix $\boldsymbol{A}_{\mathcal{U}_\mathcal{N}}$ of the projected network $\mathcal{U}_\mathcal{N}$ is the only ingredient that we need in order to characterize the topology of a network since it encodes all the information about the patterns of interaction among nodes. The number of non-zero entries of $\boldsymbol{A}_{\mathcal{U}_\mathcal{N}}$ is twice the number of links $\left|\mathcal{W}_\mathcal{N}\right| = 2L$ which cannot exceed $\binom{N}{2} = \frac{N(N-1)}{2}$. In this limit case, all possible pairs of nodes are connected by links, ergo all the off-diagonal entries of the weighted adjacency matrix $\boldsymbol{A}_{\mathcal{U}_\mathcal{N}}$ are non-zero and the network is called complete. The **link density**, ρ, is thus defined as the ratio between the actual number of links in a network and the maximum attainable one [79]:

$$\rho = \frac{2L}{N(N-1)}\,. \tag{1.3}$$

As expected for a density, $0 \leq \rho \leq 1$. At the level of single nodes, the number of links connected to a node i defines its **degree**, k_i, which is equivalent to the number of neighbors of i. Such quantity can be calculated from the weighted adjacency matrix $\boldsymbol{A}_{\mathcal{U}_\mathcal{N}}$ as the number of non-zero entries in row i. As a special case, a node with degree 0 is called **isolated**. The **average degree**, $\langle k \rangle$, of $\mathcal{U}_\mathcal{N}$ is defined as:

$$\langle k \rangle = \frac{1}{N} \sum_{j=1}^{N} k_j = \frac{2L}{N}\,, \tag{1.4}$$

where the last equality stems from the fact that a link between two nodes contributes to both degrees. The **maximum degree**, k_{max}, of $\mathcal{U}_\mathcal{N}$ is the maximum degree of its nodes:

$$k_{max} = \max_j (k_j)\,, \tag{1.5}$$

However, the degree of a node gives access only to the number of neighbors of that node, completely overlooking the intensity of the interactions. The **strength**, s_i, of a node i is the analogous of the degree which takes into account the link weights w_{ij} as

$$s_i = \sum_{j=1}^{N} w_{ij} \,. \tag{1.6}$$

Both the degree k_i and the strength s_i quantify the interactions between node i and its (first) neighbors but disregard the relations with its second neighbors. In particular, it may be the case where the neighbors of a node are likely to be connected each other. Such concept, whose origins are in sociology, goes under the name of *triadic closure* [126]. Several measures have been designed to capture it at the the level of the whole network but, historically, the first one was the **transitivity** T [89] defined as

$$T = \frac{3 \times \text{number of triangles}}{\text{number of connected triples}} \,. \tag{1.7}$$

A connected triple (also called triad) is composed by a given node and an unordered pair of links connected to it. Moreover, a triangle is a complete network of three nodes. The transitivity T can be interpreted as the fraction of transitive triples, *i.e.* those connected triples that possess the third link to form triangles. Since a transitive triple (triangle) is composed by three connected triples, it counts for each of them, thereby the factor 3 in the numerator. The transitivity is limited by the interval $0 \leq T \leq 1$, with $T \simeq 0$ possibly indicating a tree-like topology and $T \simeq 1$ an almost complete network. An alternative way to quantify the global triadic closure of a network is by means of the so-called **local clustering coefficient**, C_i, of node i defined as the ratio between the number of links, e_i, that join the neighbors of i and the maximal number of such links [127]. Since the number of neighbors of node i is equal to its degree k_i, the maximal number of links between them is $k_i(k_i-1)/2$. The local clustering coefficient is then

$$C_i = \frac{2e_i}{k_i(k_i-1)} \,, \tag{1.8}$$

a relation which holds only if $k_i > 1$, otherwise $C_i = 0$. The **average clustering coefficient**, $\langle C \rangle$, of a network is simply the average of the local clustering coefficients of the nodes in the network [127]:

$$\langle C \rangle = \frac{1}{N} \sum_{i}^{N} C_i \,. \tag{1.9}$$

Clearly, both the local and average clustering coefficients are bounded such that $0 \leq C_i \leq 1$ and $0 \leq \langle C \rangle \leq 1$. Interestingly, most real-world networks feature a remarkably high density of triangles if compared with the one of a random network with almost the same number of links and nodes, $\tilde{C} = \frac{\langle k \rangle}{N}$ [81, 82, 127]. Complementary, the large-scale structure of a network can be described using the concept of reachability of two nodes. In general, networks are not embedded in a metric

space thus a notion of distance between nodes is missing. Nevertheless, we may travel from one node to another passing through the intermediate links in the network. The number of links that we traverse is then a valid ingredient to design a reasonable equivalent of the distance. A rigorous definition of reachability is based on the notion of **path** from node i to node j, defined as an ordered set of nodes connected in sequence where the first and last elements are i and j, respectively [83]. The **length** of the path is the number of links that join the nodes i and j, namely the number of nodes in path minus one. The **shortest path** between nodes i and j is the path of minimum length between them, also referred to as **geodesic**. The **shortest path length**, l_{ij}, is then natural counterpart of the distance between nodes i and j for a network. Using the shortest path length between nodes we can gauge the typical extent of a network from the **average shortest path length** [83]

$$\langle l \rangle = \frac{1}{N(N-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} l_{ij} \,. \tag{1.10}$$

Such quantity is the average number of links in the shortest paths between every pair of nodes and measures the average degree of separation between nodes. For a random network the average shortest path length $\langle l \rangle$ scales at most logarithmically with the number of nodes N, $\langle l \rangle \sim \frac{\ln(N)}{\ln(\langle k \rangle)}$ [127–129]. Surprisingly, this same relation is valid also for real-world networks, meaning that it requires few steps to walk between any pair of nodes [130–133]. A network that exhibits such property is said to be a **small-world**. Another quantity to characterize the extent of a network is the **diameter** d, usually defined as the maximum shortest path length of a network

$$d = \max_{i,j} l_{ij} \,, \tag{1.11}$$

which specifies how far apart are the most distant nodes in the network. A whole network, however, may be composed by several parts that are eventually unrelated. Part of a graph $\mathcal{U}_\mathcal{N} = (\mathcal{V}_\mathcal{N}, \mathcal{W}_\mathcal{N})$ is formally defined as a subgraph $\mathcal{U}'_\mathcal{N} = (\mathcal{V}'_\mathcal{N}, \mathcal{W}'_\mathcal{N})$ of $\mathcal{U}_\mathcal{N}$ if all the nodes $\mathcal{V}'_\mathcal{N}$ and all the links in $\mathcal{W}'_\mathcal{N}$ are included in $\mathcal{U}_\mathcal{N}$, *i.e.* $\mathcal{V}'_\mathcal{N} \subseteq \mathcal{V}_\mathcal{N}$ and $\mathcal{W}'_\mathcal{N} \subseteq \mathcal{W}_\mathcal{N}$. A graph is called **connected** if a path exists between all pairs of nodes, otherwise it is called **unconnected** or **disconnected**. A **component** of a graph is then a connected subgraph, which is said to be a **giant component** when its size corresponds to a macroscopic fraction of the number of nodes in the graph [134]. In the case of disconnected network, a path between two nodes i and j in distinct components does not exists, thus both $\langle l \rangle$ and d are infinite. Nevertheless, in such a case these quantities are commonly redefined considering only the nodes that are part of the largest connected component [127].

Until now, we dedicated our attention to the definition of several features of a network, both at the micro- and macro-scale, by means of local and global measures, respectively. Local measures characterize nodes or links while global measures provide an overview of the network. However, we overlooked the presence of meso-scale structures at the intermediate level between the micro- and macro-scale. Such kind of structures are subgraphs that exhibit peculiar interaction patterns

among nodes. Examples of meso-scale structures include the *core-periphery* structure, which consists of a modest core of firmly connected nodes and a considerable periphery of poorly connected ones [135–137], and the *community* structure, where groups of strongly interacting nodes are observed. This second example is particularly prominent since the organization of nodes in groups is a common trait of diverse systems, ranging from biological entities [138] to social structures [139].

1.1.4 Community structure

Communities, also known as **clusters** or **modules**, are described as subgroups of nodes where connections are tighter within them than with the rest of nodes [140]. Given a network, a **partition** is then defined as a specific arrangement of nodes into subgroups. Since there are many possible partitions of a network, we need to assess which among them contain appropriate communities that match the above definition. To this aim, several *quality functions* have been introduced, each one providing a different criterion to quantify if a given partition exhibits a real community structure or not [140–142]. One of the most popular quality functions is the **modularity**, introduced by Newman and Girvan in their seminal paper [143] and further analyzed in [144]. The definition of modularity is simply based on a reasonable assumption: a random network does not feature a community structure. In this case, a random network is devised by reshuffling the links of the original network while preserving the degree of each node as defined in subsection 1.1.3. Given two nodes i and j with degree k_i and k_j, respectively, the probability that they are connected in a random network is then:

$$p_{ij} = \frac{k_i k_j}{2L}, \tag{1.12}$$

where L is the total number of links in the network. Given a partition of the original network, we thus take a group of nodes comparing its link density with the one calculated for the same group in a random network. If the link density in the original network is greater than expected in the random case, we can conclude that the interaction pattern inside the group is distinct from the random one and the group itself is a community. Moreover, the larger the difference between expected and observed densities, the higher the probability that the group is a community.

Repeating the same procedure for every group of nodes in a given partition, we evaluate the overall quality of the partition using the modularity Q defined as:

$$Q = \frac{1}{2L} \sum_{i,j=1}^{N} \left(A_{ij} - \frac{k_i k_j}{2L} \right) \delta\left(C_i, C_j\right), \tag{1.13}$$

where N is the number of nodes in the network and $\boldsymbol{A}$ is the adjacency matrix. The term $\frac{k_i k_j}{2L}$ is the expected link density in a random network, (Equation 1.12). C_i and C_j are the groups to which nodes i and j belong to, respectively, and the sum is only valid for pairs of distinct nodes (thus $i \neq j$). Finally, the Kronecker delta, $\delta\left(C_i, C_j\right)$ is 1 if nodes i and j are in the same group,

i.e. $C_i = C_j$, otherwise is 0. Due its presence, the only contributions to the sum in the modularity Q come from nodes in the same group. Hence, Equation 1.13 can be recast as the sum over the groups

$$Q = \sum_{m=1}^{n_g} \left[\frac{L_m}{L} - \left(\frac{K_m}{2L} \right)^2 \right], \tag{1.14}$$

where n_g is the number of groups, L_m is the number of links inside group m and K_m is the sum of the degree of the nodes in the group. Therefore, for a fixed group m, the first term in Equation 1.14 represents the fraction of links of the network that join the nodes in group m, while the second term is the expected fraction of links within group m for a random network where the degree of every node is the same as in the original network. Overall, summing all the contributions from each group m, high positive values of the modularity Q denotes a partition with a well-defined community structure. In particular, the modularity is always smaller than one and attains zero when the entire network is considered as a single group since the only two terms that contribute to the modularity are identical and they cancel out each other. On the other hand, modularity can have negative values, *e.g.* in the case of a partition where each node is a separate group. Therefore, if no partitions of a network have positive modularity the network itself does not possess a community structure. We remark that the modularity Q in Equation 1.14 is not only an explicit formulation which encapsulates the definition of communities as dense subgraphs of nodes, but it has been derived also in the well-founded framework of maximum likelihood approaches to community detection [145]. Despite the good properties of the modularity, several drawbacks have been discovered. For example, modularity is affected by a resolution limit that produces aggregated communities despite they are clearly separated in a simple synthetic network [146]. Another, more general limitation is due to the fact that the modularity landscape of real networks is often degenerate, in the sense that many partitions of a network have modularity values which correspond to local maxima very close to each other. In addition, these maxima forms a plateau which is close to the global maximum of the modularity. As a consequence, networks may not admit a precise maximum of the modularity [147]. Last but not least, random networks may comprise partitions with large modularity values, in contrast to the hypothesis that suggested the formulation of modularity [148–150].

The modularity Q, devised for unweighted networks, can be easily extended to the weighted case. Taking as a reference the formulation of modularity in Equation 1.13, the adjacency matrix $\boldsymbol{A}$ is replaced with the weight matrix $\boldsymbol{W}$ and the degrees k_i and k_j are substituted with the strengths s_i and s_j of nodes i and j, defined in Equation 1.6. Consistently, the total number of links L is replaced by the sum of the link weights W in order to ensure that the expected probability of connection between nodes i and j in a random network is properly transformed into the expected link weight p_{ij}^w between the same nodes in a weighted random network where the strength of nodes has been preserved:

$$p_{ij}^w = \frac{s_i s_j}{2W}. \tag{1.15}$$

The **weighted modularity** Q^w can be derived from Equation 1.13, using the abovementioned substitutions, as

$$Q^w = \frac{1}{2W} \sum_{i,j=1}^{N} \left(W_{ij} - \frac{s_i s_j}{2W} \right) \delta\left(C_i, C_j\right). \tag{1.16}$$

In turn, this equation can be rearranged in the same form of Equation 1.14:

$$Q^w = \sum_{m=1}^{n_g} \left[\frac{W_m}{W} - \left(\frac{S_m}{2W} \right)^2 \right], \tag{1.17}$$

where n_g is the number of groups, W_m is the sum of the weights of links inside group m and S_m is the sum of the strengths of the nodes in the group. For a fixed group m, the first and second term in Equation 1.17 are, respectively, the observed and expected fraction of the total weight of the network inside group m, where the latter is calculated for the random counterpart of the original network in which the strength of each node is unchanged but the link wights are randomly assigned.

Thanks to the introduction of modularity, we are now able to evaluate the quality of any partition of a network in a quantitative fashion. The goal of this operation, indeed, is to find the partition with the best community structure, *i.e.* with the highest modularity. In principle, we should compute the modularity for every partition of the network. However, the number of possible partitions grows faster than exponentially with the number of nodes in the network [80, 140, 151]. As a consequence, a systematic check of the quality of all the partitions is not computationally viable. Therefore, the best we can do is to design algorithms that try to discover good partitions by maximizing modularity. A broad class of these algorithms is based on greedy optimization [152], a heuristic technique that always performs, at each step, the best local choice towards the optimization of a score function. Nonetheless, there is no guarantee that the globally optimal solution will be reached trough intermediate steps which are locally beneficial. Indeed, in most of the cases, greedy strategies may achieve locally optimal solutions that are not far from the optimal one.

In the context of modularity maximization, a well-know greedy method is the *Louvain algorithm* [153]. Each step of the algorithm is divided in two stages:

Stage I At the beginning, each node of a weighted network is assigned to a different community, thus $n_g = N$. For each node i, we examine its neighbors j calculating the change in modularity that would occur if we remove node i from its community and place it in the community of j. Among all the possible placements, we pick the one which produces the highest increase of modularity, moving the node i to the corresponding community. Otherwise, if no increase is observed, i remains in its original community. This elementary operation is repeated sequentially for all nodes until no individual move can increase the modularity. In such case, a local maximum of the modularity is reached so the first stage is completed. We notice that a node may be considered several times during the stage,

as it is often the case. The computational advantage of the algorithm stems from the fact that the change in modularity can be calculated in a close form that only requires local information about the connections of the node that we move and the target community. To be more precise, the modularity change, ΔQ_m, obtained after moving a given node i into the community m of its neighbor j is:

$$\Delta Q_m = \left[\frac{W_m + w_{i,in}}{2W} - \left(\frac{S_m + s_i}{2W} \right)^2 \right] - \left[\frac{W_m}{2W} - \left(\frac{S_m}{2W} \right)^2 - \left(\frac{s_i}{2W} \right)^2 \right], \tag{1.18}$$

where W_m is the sum of the weights of the links that join nodes in community m, $w_{i,in}$ is the sum of the weights of the links between i and all the nodes in community m, S_m is the sum of the strength of nodes in community m, s_i is the strength of node i and W is the sum of the link weights in the network. In order to move node i into a community m, we have to remove it beforehand from the community C_i it belongs to. The change in modularity for doing so is then $-\Delta Q_{C_i}$. Overall, one has to compute first the change in modularity by removing node i from its community C_i and then by moving it into a community m of one of its neighbors.

Stage II The communities identified in the previous stage becomes the nodes of a new network. The weight of the link between two nodes is constructed as the sum of the link weights between nodes in the corresponding communities. Moreover, nodes have self-loops whose weights are the sum of the link weights between nodes in the same community. Once the construction of the new network is completed, the second stage terminates.

Once a step composed by the two stages is completed, the algorithm iterates over the steps until no further improvement in the modularity is possible, meaning that the maximum value of the modularity has been reached. Most of the computations are performed in the first steps of the algorithm since the number of communities decreases rapidly at each step. Typically, few steps are sufficient to reach the maximum of modularity. By construction, the final output of the algorithm is a hierarchy of communities where the top level is the one with the highest modularity. However, the outcome of any greedy method depends on the initial condition. In the present case, the sequence in which nodes are parsed influences the final partition. Therefore, the Louvain algorithm is usually applied several times on a network, parsing the nodes at random in each run. The partition with the best modularity among all the runs is then retained.

The Louvain algorithm became popular among the community detection methods thanks to the good quality of the resulting partition and the competitive performance in terms of computational resources. Besides greedy techniques, other strategies have been designed to optimize modularity [140]. Moreover, the literature on community detection methods have been flourishing over the years comprising different quality functions as well as several techniques to optimize them [140, 154]. Recent developments are more focused on a data-driven approach to discover latent communities, somehow relaxing the definition of quality function towards an implicit description of the communities [155]. This perspective has been introduced by computer scientists,

among which the community detection problem is a very active research field. Some of the most distinguished contributions are *Infomap* [156] and the stochastic block model [157]. In particular, the former is based on encoding the trajectory of a random walker on a network in the most efficient way [156, 158, 159]. In this case, the greedy optimization strategy of the Louvain algorithm is applied to minimize the map equation, a quality function that measures the compression of the trajectory based on the entropy (section 1.2). Finally, various generalizations of the community structure have been introduced in order to highlight different flavors of meso-scale patterns within a single framework [160, 161].

1.1.5 Measures of similarity between partitions

As we have seen above, detecting communities is a task that can be accomplished in several ways. Then, a question arises: how can we compare partitions uncovered by different methods? Similarity measures provide a rigorous tool to assess the resemblance of two partitions, even in the case where one of them correspond to a **ground-truth** classification of nodes. Specifically, depending on the entities we want to compare, similarity measures can be divided in two main categories. The first includes measures to evaluate the pairwise correspondence between selected communities, while the second consist of measures that indicate the overall similarity between two partitions as the result of the comparison between the respective communities.

The first type of indicators are commonly based on counting the number of nodes that are present in two communities, say C and D, each composed by a given set of nodes. The number of shared nodes, n_{CD}, that belong to both communities is the cardinality of the intersection between the two sets of nodes C and D:

$$n_{CD} = \left|C \cap D\right|. \tag{1.19}$$

As the size of the intersection between C and D is affected by the size of the two communities, we should take into account their size to properly compare them. For example, n_{CD} can be divided by the size of the maximum or minimum between the two sets, C and D, or by a combination of their sizes like the sum or the product. A possible option is to divide it by the size of the union of the two sets, $\left|C \cup D\right|$, which leads to the **Jaccard score** J_{CD} [162, 163] given by

$$J_{CD} = \frac{n_{CD}}{\left|C \cup D\right|} = \frac{\left|C \cap D\right|}{\left|C \cup D\right|} = \frac{\left|C \cap D\right|}{\left|C\right| + \left|D\right| - \left|C \cap D\right|}. \tag{1.20}$$

The Jaccard score, J_{CD}, turns out to be a thoughtful choice for measuring the relative size of the overlap between two sets since it has a simple heuristic interpretation [164]: it quantifies the probability that an element belonging to at least one of the two sets (*i.e.* included in the union) is also an element of both of them. Furthermore, it is a metric defined on finite sets [164, 165] which allows to properly compare the similarities calculated for different communities. Finally, the Jaccard score J_{CD} varies between 0 and 1, where the former denotes an empty intersection between C and D (meaning that the two sets are totally different), while the latter indicates that

C and D are identical.

The second type of indicators quantify the correspondence between partitions. Since the partition of a network is composed of several communities, some of the measures introduced before may be adopted also as basic ingredients to describe the similarity between partitions. Various definitions are possible but the most widespread are all grounded in Information Theory [166] which provides a powerful framework for their derivation. The idea behind these measures is that the more similar the partitions are, the least information we need to reconstruct one partition from the other. The amount of such information is then a measure of their dissimilarity. Given a network composed by N nodes which are divided in two partitions $\mathcal{X} = \{x_i\}_{i=1}^{n^{\mathcal{X}}}$ and $\mathcal{Y} = \{y_j\}_{j=1}^{n^{\mathcal{Y}}}$; the elements x_i and y_j of each partition denotes the sets of nodes that correspond to the communities in partitions $\mathcal{X}$ and $\mathcal{Y}$, respectively. In principle, every node can be the member of any community in both partitions $\mathcal{X}$ and $\mathcal{Y}$. In probabilistic terms, we can imagine that the community memberships of a node in each partition, *i.e.* x_i and y_j, are the values of two random variables X and Y [140, 166]. The probability that a node picked at random is present both in community $x_i \in \mathcal{X}$ and $y_j \in \mathcal{Y}$ is then the joint probability distribution $P(x_i, y_j)$ of the two random variables X and Y:

$$P\big(x_i, y_j\big) = P\big(X = x_i, Y = y_j\big) = \frac{n_{ij}}{N}\,, \tag{1.21}$$

where n_{ij} is the same quantity in Equation 1.19. The marginal probabilities, obtained by consistence, are

$$\begin{aligned} P\big(x_i\big) &= \sum_{j=1}^{n^{\mathcal{Y}}} P\big(X = x_i, Y = y_j\big) = \frac{n_i^X}{N} \\ P\big(y_j\big) &= \sum_{i=1}^{n^{\mathcal{X}}} P\big(X = x_i, Y = y_j\big) = \frac{n_j^Y}{N}\,. \end{aligned} \tag{1.22}$$

where n_i^X and n_j^Y are the size of the communities x_i and y_j, respectively. The **mutual information**, $\mathrm{MI}(X, Y)$, quantifies the degree of dependence between two random variables X and Y, measuring the average decrease in uncertainty about one variable after knowing the other one. Since the partitions $\mathcal{X}$ and $\mathcal{Y}$ can be characterized themselves by random variables X and Y, the mutual information $\mathrm{MI}\big(\mathcal{X}, \mathcal{Y}\big)$ between two partitions $\mathcal{X}$ and $\mathcal{Y}$ is naturally defined as the mutual information between the random variables X and Y associated to the partitions $\mathcal{X}$ and $\mathcal{Y}$, respectively

$$\mathrm{MI}\big(\mathcal{X}, \mathcal{Y}\big) = \sum_{i=1}^{n^{\mathcal{X}}} \sum_{j=1}^{n^{\mathcal{Y}}} P\left(x_i, y_j\right) \ln \frac{P\left(x_i, y_j\right)}{P\big(x_i\big) P\big(y_j\big)}\,. \tag{1.23}$$

The mutual information is non-negative, $(\mathrm{MI}(\mathcal{X}, \mathcal{Y}) \geq 0)$; it is symmetric, $(\mathrm{MI}(\mathcal{X}, \mathcal{Y}) = \mathrm{MI}(\mathcal{Y}, \mathcal{X}))$; and $\mathrm{MI}\big(\mathcal{X}, \mathcal{Y}\big) = 0$ if and only if the two random variables $\mathcal{X}$ and $\mathcal{Y}$ are independently distributed, *i.e.* $P(x_i, y_j) = P(x_i)P(y_j)$. The definition of the mutual information as a measure of dependence between two random variables is then justified, since in the case of independent random variables

knowing one of the two variables does not provide any information about the other, *i.e.* its uncertainty is not reduced. In order to better appreciate the meaning of mutual information $\mathrm{MI}(\mathcal{X},\mathcal{Y})$ as the reduction in uncertainty about *e.g.* $\mathcal{X}$, Equation 1.23 can be rewritten as

$$\mathrm{MI}(\mathcal{X},\mathcal{Y}) = \mathrm{H}(\mathcal{X}) - \mathrm{H}(\mathcal{X}|\mathcal{Y}), \tag{1.24}$$

where $\mathrm{H}(\mathcal{X})$ is the entropy of $\mathcal{X}$

$$\mathrm{H}(\mathcal{X}) = -\sum_{i=1}^{n^{\mathcal{X}}} P(x_i)\ln P(x_i), \tag{1.25}$$

which represents the **uncertainty**[1] of $\mathcal{X}$ and $\mathrm{H}(\mathcal{Y}|\mathcal{X})$ is the conditional entropy of $\mathcal{X}$ given $\mathcal{Y}$, *i.e.* the uncertainty about $\mathcal{X}$ after observing $\mathcal{Y}$

$$\mathrm{H}(\mathcal{Y}|\mathcal{X}) = \sum_{i=1}^{n^{\mathcal{X}}}\sum_{j=1}^{n^{\mathcal{Y}}} P(x_i,y_j)\ln P(x_i|y_j). \tag{1.26}$$

Therefore, for independent partitions $\mathrm{H}(\mathcal{X}|\mathcal{Y}) = \mathrm{H}(\mathcal{X})$, which implies $\mathrm{MI}(\mathcal{X},\mathcal{Y}) = 0$. Unfortunately, the mutual information is not the best similarity measure between partitions that can be designed: indeed, any partition $\mathcal{X}'$ obtained from $\mathcal{X}$ by further dividing its communities has the same mutual information with $\mathcal{X}$, $\mathrm{MI}(\mathcal{X},\mathcal{X}') = \mathrm{H}(\mathcal{X})$, since the conditional entropy is systematically zero, $\mathrm{H}(\mathcal{X}|\mathcal{X}') = 0$. To overcome such limitation, the **normalized mutual information** NMI [167] has been defined

$$\mathrm{NMI}(\mathcal{X},\mathcal{Y}) = \frac{2\,\mathrm{MI}(\mathcal{X},\mathcal{Y})}{\mathrm{H}(\mathcal{X})+\mathrm{H}(\mathcal{Y})}. \tag{1.27}$$

which is also symmetric and ranges between 0 and 1. Indeed, NMI is equal to 0 if and only if $\mathcal{X}$ and $\mathcal{Y}$ are independent and is equal to 1 only when the partitions are identical, $\mathcal{X} = \mathcal{Y}$. Another variant of the mutual information MI is the **variation of information** VI [168] which measures the amount of information lost and gained when changing from partition $\mathcal{X}$ to partition $\mathcal{Y}$:

$$\mathrm{VI}(\mathcal{X},\mathcal{Y}) = \mathrm{H}(\mathcal{X}) + \mathrm{H}(\mathcal{Y}) - 2\,\mathrm{MI}(\mathcal{X},\mathcal{Y}) = \mathrm{H}(\mathcal{X},\mathcal{Y}) - \mathrm{MI}(\mathcal{X},\mathcal{Y}) = \mathrm{H}(\mathcal{X}|\mathcal{Y}) + \mathrm{H}(\mathcal{Y}|\mathcal{X}). \tag{1.28}$$

Remarkably, the variation of information VI is a distance in the space of partitions since it is symmetric, non-negative and fulfills the triangle inequality. The last property implies that the variation of information is upper bounded, *i.e.* $\mathrm{VI}(\mathcal{X},\mathcal{Y}) \leq \ln N$, where the maximum value is achieved, for example, if one partition is composed by a single community of N nodes, and the other by as many communities as nodes. Since the maximum value of the VI depends on the number of nodes N, VI cannot be directly adopted to compare networks of different size. However, simply dividing VI by its maximum, $\ln N$, provides a normalized measure that can be

[1] For the sake of readability, we do not provide here a justification for the interpretation of entropy as a measure of uncertainty. We point the interested reader to the next Section, where we detail a thorough derivation of entropy and we fully motivate this interpretation.

employed for such comparisons. As a direct consequence of its metric properties, VI is a local measure which depends on the discrepancies between two partitions only in the portion of the network where they take place and not on the partition of the rest of the graph [168].

In conclusion, NMI and VI are interdependent

$$\mathrm{NMI}(\mathcal{X},\mathcal{Y}) = 1 - \frac{\mathrm{VI}(\mathcal{X},\mathcal{Y})}{\mathrm{H}(\mathcal{X}) + \mathrm{H}(\mathcal{Y})},$$

and both have their own appealing properties. NMI quantifies the similarity between partitions through their mutual dependence, with minimum and maximum values corresponding to well-defined characteristics of the partitions. On the other hand, VI measures the distance between partitions and different values can be properly compared. For example, given two partitions $\mathcal{Y}'$ and $\mathcal{Y}''$, if their similarities with a third partition $\mathcal{X}$ are such that $\mathrm{VI}(\mathcal{Y}',\mathcal{X}) = 2\,\mathrm{VI}(\mathcal{Y}'',\mathcal{X})$ then $\mathcal{Y}'$ and $\mathcal{X}$ are two times closer than $\mathcal{Y}''$ and $\mathcal{X}$. However, the same is not valid for NMI: a value twice as much as another, like $\mathrm{NMI}(\mathcal{Y}',\mathcal{X}) = 2\,\mathrm{NMI}(\mathcal{Y}'',\mathcal{X})$, does not imply that $\mathcal{Y}'$ and $\mathcal{X}$ are two times more dependent than $\mathcal{Y}''$ and $\mathcal{X}$.

However, existing measures to compare partitions are biased, for example, on the number of clusters [168], the cluster size distribution [169], or they may not properly account random partitions [170]. Moreover, none of the measures easily generalize to overlapping and hierarchical partitions. To overcome such biases and limitations, a new measure has been proposed that applies to disjoint, overlapping and hierarchical partitions [171]. Such measure shifts the problem of comparing partitions from a cluster-centric to an element-centric perspective, where the partition induces relationships between elements that are used to compare them. As a consequence, the measure follows closely the behavior expected from intuition.

We already mentioned that the entropy is a measure of **uncertainty**. Nonetheless, we did not derived it from first principles nor we stressed enough its importance. The next Section is then devoted to properly introduce the notion of entropy and clarify its fundamental meaning, along with its properties.

1.1.6 Filtering methods for weighted networks

A complex system represented as a weighted network usually exhibits patterns of interactions that are enriched by the presence of the connection weights. Despite such weights are an additional source of information that allows to model more realistically many systems, they may encode relationships that are not significant. Therefore, it may happens that a substantial fraction of the links does not carry important information. Over the years, several criteria to prune irrelevant connections have been developed [172–175]. Clearly, if we have some knowledge about the mechanism that assigns the weights, we can devise a method that establish if a given weight is significant based on the expected value under a null model where, *e.g.* the process generating the weights has some random characteristic. However, in general, the mechanism that is responsible

for the creation of the weights is unknown or we cannot design a reasonable random expectation of the weight values. In these situations, different methods have been introduced to establish the significance of a weight from various statistical features of the topology of the original network, all of them performing the filtering in an *ex-post* way, *i.e.* after the computation of the link weight.

An example of a local method is the disparity filter described in [172] where, for each node, every link weight is compared to the strength of the node. Only the links whose weight is significantly higher that the weight expected from a random null model that satisfies the constraint on the strength are retained. In this way, if the weights of the node are homogeneously distributed no link is preserved for the current node. As a result, the links of the network that are conserved are the ones that connect two nodes and are significant for at least one of them. Despite drastically reducing the network density, the disparity filter retains nearly all the nodes and preserves both the degree and weight distribution along with the clustering coefficient. However, such local filtering may be not suited for every real-world network. Indeed, we may be interested in the conservation of salient features of the network topology based on a global criterion. In such a case, another method has been designed to preserve the weight distribution from a null model that consider all the weights [173].

Nevertheless, both methods described above do not guarantee that the inferred pattern of connections is essential in the sense that it cannot be further reduced while conserving the same properties of the network. A scrupulous approach to achieve such irreducibility relies on formulating null models via the maximum entropy principle [176]. Following this prescription, only the links whose properties cannot be reconstructed from local information at the node level are then retained. Different null models are then possible depending on the constraints that we want to preserve as ensemble averages, namely the degree of the nodes [177], the strength [178, 179] or both of them [174, 180].

1.2 Entropy

Historically, the concept of entropy dates back to 1865 when it was introduced by Rudolf Clausius to describe the macroscopic behavior of a thermodynamic system at the equilibrium. Thanks to its definition, it was possible to reformulate the second law of thermodynamics [181]. In 1870, Ludwig Boltzmann restated the definition of entropy in terms of the microscopic states of a system at the equilibrium, providing the link between its microscopic features and the macroscopic realm [182]. This seminal formulation laid the foundation of *statistical mechanics*, a branch of Physics which describes the behavior of interacting systems trough the statistical properties of their constituents. Later on, Josiah W. Gibbs generalized the expression of the Boltzmann entropy to characterize a system in terms of the probability of its microstates. An analogous definition was then introduced by Claude Shannon in a landmark paper about Information Theory dated 1948 with the purpose of studying the efficiency of communication of a message [183].

To understand the notion of **information entropy**, consider an event X which has n possible

outcomes $\{x_1, x_2, \ldots, x_n\}$. A message specifying the actual outcome, say x_i, is transmitted to a receiver when the event happens. Information is effectively present in the message only if the receiver does not know *a priori* (with certainty) the content of the message, otherwise the information content of the message is null. Before the communication, there is an uncertainty about the occurrence of x_i which disappears after, since the information in the message arrived to the receiver. The acquisition of information for the receiver, indeed, cancels the uncertainty. The **self-information** $\mathrm{I}(x_i)$ [184] is then defined as the *a priori* uncertainty regarding the occurrence of the outcome x_i, which depends only on its occurrence probability $P(x_i)$ and not on the value of the outcome itself x_i:

$$\mathrm{I}(x_i) = f(P(x_i)) . \tag{1.29}$$

In order to determine the form of function f we require it to satisfy the following properties:

- Self-information is a continuous function of $P(x_i)$ and decreases as $P(x_i)$ increases.
- Self-information is non-negative, *i.e.* $\mathrm{I}(x_i) \geq 0$, and the equality holds only if the outcome x_i is deterministic, *i.e.* $P(x_i) = 1$.
- Self-information is additive: if x_i is composed by two independent outcomes x_{i1} and x_{i2} then $x_i = x_{i1} \cap x_{i2}$; hence, the additive property implies:

$$\mathrm{I}(x_i) = \mathrm{I}(x_{i1}) + \mathrm{I}(x_{i2}) , \tag{1.30}$$

$$f(P(x_i)) = f(P(x_{i1})) + f(P(x_{i2})) . \tag{1.31}$$

  This relation simply tells us that the two outcomes do not influence each other, therefore the joint self-information is decoupled. The independence of the outcomes, in terms of the probabilities of occurrence, reads:

$$P(x_{i1} \cap x_{i2}) = P(x_{i1})\, P(x_{i2}) ,$$

  which plugged in Equation 1.31 leads to

$$f(P(x_{i1})\, P(x_{i2})) = f(P(x_{i1})) + f(P(x_{i2})) . \tag{1.32}$$

The solution to the functional equation (1.32) is then a functional f which satisfies the above conditions. The only admitted functional corresponds to the logarithm. Since the basis of the logarithm must be specified, the more general solution includes a multiplicative constant that incorporates an arbitrary change of basis:

$$f(P(x_i)) = K \log(P(x_i)) . \tag{1.33}$$

Since the probability, by definition, ranges between 0 and 1 and the self-information $\mathrm{I}(x_i)$ is always non-negative, the constant $K < 0$. Thus, without loosing the generality, we can set $K = -1$

remembering that any base for the logarithm is legitimate. The only influence of the base of the logarithm is on the units of $\mathrm{I}(x_i)$: *e.g.* , if the logarithm is to the base 2, then $\mathrm{I}(x_i)$ is expressed in bits, where $\log_2 \frac{1}{2} = 1\text{bit}$. The name *bit* stems from the contraction of *binary digit*, adopted for the first time in the Shannon's milestone paper [183] as the basic unit of information in computing and communication theory. Another common option for the base is the Napier number (also called Euler constant) e, for which $\log_e = \ln$; in analogy with the *bit*, the *nat* is then defined as the natural unit of information entropy, as $\ln \frac{1}{e} = 1\text{nat}$. This definition represents the choice of 1 from e and is the common one in statistical mechanics for the Gibbs entropy. Finally, the self-information $\mathrm{I}(x_i)$ of the outcome x_i with associated probability $P(x_i)$ reads:

$$\mathrm{I}(x_i) = -\ln\big(P(x_i)\big) = \ln\left(\frac{1}{P(x_i)}\right). \tag{1.34}$$

A small probability of the outcome x_i entails a large uncertainty associated to the occurrence of x_i, meaning that the self-information $\mathrm{I}(x_i)$ carried by the message after the outcome x_i actually happened is high.

Every possible outcome $\{x_1, x_2, \ldots, x_n\}$ of an event X has a self-information $\mathrm{I}(x_i)$ as described in Equation 1.34. The event X can then be represented as a discrete random variable with probability mass function $\mathbf{P}(X)$ which specifies the probability of each outcome, *i.e.* the probability that the random variable is equal to a given value $P(X = x_i) = P(x_i) \geq 0$. Moreover, by definition, the probability mass function is normalized, $\sum_{i=1}^{n} P(x_i) = 1$. The (information) entropy $S[X]$ of a discrete random variable X is the expected value of the self-information of its outcomes or, in other words, the average self-information per outcome:

$$S[X] = \mathbf{E_P}[\mathrm{I}(X)] = \mathbf{E_P}[-\ln(\mathbf{P}(X))] = \sum_{i=1}^{n} P(x_i)\,\mathrm{I}(x_i) = -\sum_{i=1}^{n} P(x_i)\ln P(x_i), \tag{1.35}$$

Alternatively, it can be interpreted as the average uncertainty associated *a priori* to the content of a message. The information entropy is a continuous function in each variable $P(x_i)$ since it is the sum of continuous functions. Moreover, the information entropy is additive since it is the average value of an additive quantity(*i.e.* the self-information) over the possible outcomes. In the case when $P(x_i) = 0$ for some i, we define the term $P(x_i)\ln P(x_i)$ to be 0, since the limit $\lim_{P(x_i)\to 0^+} P(x_i)\ln P(x_i) = 0$. The same result holds when $P(x_i) = 1$, thus we obtain an entropy $S[X] = 0$ when the result of the event X is deterministic, *i.e.* $P(x_i) = 1$ for a fixed i and $P\big(x_j\big) = 0$ for every other $j \neq i$. The uncertainty about the result of the random variable X, as expressed by $S[X]$, is then null since the outcome is always x_i. On the contrary, if $P(x_i) = \frac{1}{n}$ for every i, the entropy attains its maximum value, $S_{max}[X] = \ln(n)$: the average uncertainty about the presence of an outcome in the message is maximum when all the outcomes are equally probable. The maximum entropy $S_{max}[X]$ increases monotonically with the number of outcomes: given two discrete random variables X_1 and X_2 with n_1 and n_2 possible outcomes respectively, if $n_1 > n_2$ then $S_{max}[X_1] > S_{max}[X_2]$. Intuitively, when there are more possible outcomes which are equally likely there is more choice about the one to send in the message, *i.e.* the average uncertainty about the received outcome is greater.

In many practical scenarios, we may be interested in comparing the observed probability distribution $\mathbf{P}(X)$ that describes an event X with a model $\mathbf{Q}(X)$ for the expected occurrence of that event. A proper measure to quantify the difference between two probability mass functions $\mathbf{P}(X)$ and $\mathbf{Q}(X)$ is the **relative entropy** of $\mathbf{P}(X)$ with respect to $\mathbf{Q}(X)$, commonly known as the **Kullback-Leibler divergence** between $\mathbf{P}(X)$ and $\mathbf{Q}(X)$ [185, 186]:

$$\begin{aligned} D_{\mathrm{KL}}(\mathbf{P}||\mathbf{Q}) = \mathbf{E}_{\mathbf{P}}\left[\ln\left(\frac{\mathbf{P}(X)}{\mathbf{Q}(X)}\right)\right] = \sum_{k=1}^{n} P(x_k)\ln\left(\frac{P(x_k)}{Q(x_k)}\right) = \\ = -\sum_{k=1}^{n} P(x_k)\ln Q(x_k) + \sum_{k=1}^{n} P(x_k)\ln P(x_k)\,. \end{aligned} \tag{1.36}$$

The Kullback-Leibler divergence is the difference between the expected values of the self-information computed according to the model $\mathbf{Q}$ and the observed distribution $\mathbf{P}$, respectively, where both expectations are taken using the probability $\mathbf{P}$. The Kullback-Leibler divergence quantifies the additional uncertainty that is present if one assumes that the distribution of X is $\mathbf{Q}$ instead of $\mathbf{P}$ [187]. This interpretation is evident from the last equality in Equation 1.36 as the last summand is the opposite of the information entropy in Equation 1.35. The Kullback-Leibler divergence is not a distance between probability distributions since it is not symmetric and does not satisfy the triangle inequality. However, $D_{\mathrm{KL}}(\mathbf{P}||\mathbf{Q})$ is always non-negative and is zero if and only if $\mathbf{P} = \mathbf{Q}$. In the definition of Equation 1.36 if $\mathbf{P}(x_i) = 0$ for some x_i then, by continuity, $0\ln\frac{0}{Q(x_i)} = 0$ since the corresponding limit is zero. By convention, when both probabilities are zero $0\ln\frac{0}{0} = 0$ but, if only $\mathbf{Q}(x_i) = 0$ for some x_i then $\mathbf{P}(x_i)\ln\frac{\mathbf{P}(x_i)}{0} = \infty$ and the Kullback-Leibler divergence is not defined. The mutual information $\mathrm{MI}\big(X,Y\big)$, introduced in the previous subsection 1.1.5, Equation 1.23, is defined as the relative entropy $D_{\mathrm{KL}}\big(\mathbf{P}(X,Y)||\mathbf{P}\big(X\big)\mathbf{P}\big(Y\big)\big)$ of the observed joint probability $\mathbf{P}(X,Y)$ with respect to the model probability $\mathbf{P}\big(X\big)\mathbf{P}\big(Y\big)$ where the events X and Y are independent, as given in Equation 1.22:

$$\begin{aligned} \mathrm{MI}\big(X,Y\big) = D_{\mathrm{KL}}\big(\mathbf{P}(X,Y)||\mathbf{P}\big(X\big)\mathbf{P}\big(Y\big)\big) = \mathbf{E}_{\mathbf{P}(X,Y)}\left[\ln\left(\frac{\mathbf{P}(X,Y)}{\mathbf{P}(X)\,\mathbf{P}(Y)}\right)\right] = \\ = \sum_{i=1}^{n^X}\sum_{j=1}^{n^Y} P\big(x_i,y_j\big)\ln\left(\frac{P\big(x_i,y_j\big)}{P\big(x_i\big)P\big(y_j\big)}\right). \end{aligned} \tag{1.37}$$

Information entropy, however, is not the only measure of uncertainty that have been designed. Indeed, several alternatives have been introduced, but the most immediate generalization of the information entropy is the Rényi entropy [188] which includes information entropy and other simpler definitions of entropy as special cases. Rényi entropy fulfills the same properties of information entropy, namely, is a continuous, non-negative function of the probabilities $P(x_i)$, it has the same maximum of information entropy for equiprobable outcomes and is additive. A more general formulation of entropy is the Tsallis entropy [189] which breaks the additivity of information entropy. Tsallis entropy was proposed as a suitable quantity to characterize out-of-equilibrium systems that do not obey the Boltzmann-Gibbs theory of statistical mechanics but are described by non-extensive statistics [190, 191].

The notion of information entropy can be extended to random variables which take continuous values. In such case, a continuous random variable X is described by a probability density function $f(x)$ which satisfies the normalization condition $\int_{-\infty}^{+\infty} f(x)\,dx = 1$. The analogue of information entropy is the **differential entropy** $s[X]$, also known as continuous entropy:

$$s[X] = \mathbf{E}_f\left[-\ln f(x)\right] = -\int_{-\infty}^{+\infty} f(x)\ln f(x)\,dx. \tag{1.38}$$

However, the naive substitution of the sum in Equation 1.35 with the integral induces a problem: although $f(x) > 0$, there is no guarantee that $f(x) \leq 1\ \forall x$. If $\exists\,(a,b) : f(x) > 1$ for $x \in (a,b)$ the contribution of that interval to the differential entropy is negative (since $\ln f(x) > 0$) and such negative part may not be counterbalanced by other positive contributions. Hence, the differential entropy can be negative, unlike the discrete (information) one. In order to understand more in depth the relation between the differential and information entropies we start by discretizing the former. In this way, we should obtain an entropy that resembles the discrete one. Thus, consider that we split X into bins of width Δ. If the probability density function is continuous inside the bins, according to the mean value theorem, there exists a value x_i inside each bin such that the associated probability p_i is

$$p_i = f(x_i)\,\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)\,dx, \tag{1.39}$$

from which follows that the discretized version X^Δ of the continuous random variable X is simply

$$X^\Delta = x_i \qquad \text{if} \quad i\Delta \leq X \leq (i+1)\Delta.$$

The entropy of the discretized random variable X^Δ is

$$\begin{aligned} S\left[X^\Delta\right] &= -\sum_{i=-\infty}^{+\infty} p_i \ln p_i \\ &= -\sum_{i=-\infty}^{+\infty} f(x_i)\,\Delta \ln\left(f(x_i)\,\Delta\right) = -\sum_{i=-\infty}^{+\infty} f(x_i)\,\Delta \ln\left(f(x_i)\right) - \sum_{i=-\infty}^{+\infty} f(x_i)\,\Delta \ln(\Delta)\,. \end{aligned} \tag{1.40}$$

If both terms $f(x)\ln\left(f(x)\right)$ and $f(x)$ in $S\left[X^\Delta\right]$ are Riemann integrable, by definition of Riemann integrability we have

$$\lim_{\Delta\to 0} \sum_{i=-\infty}^{+\infty} f(x_i)\,\Delta \ln\left(f(x_i)\right) \to \int_{-\infty}^{+\infty} f(x)\ln f(x)\,dx = s[X]\,,$$

$$\lim_{\Delta\to 0} \sum_{i=-\infty}^{+\infty} f(x_i)\,\Delta \to \int_{-\infty}^{+\infty} f(x)\,dx = 1\,.$$

Using these relations in Equation 1.40 we obtain for the differential entropy the following expression

$$s[X] = \lim_{\Delta\to 0}\left(S\left[X^\Delta\right] + \ln\Delta\right) \to -\int_{-\infty}^{+\infty} f(x)\ln f(x)\,dx. \tag{1.41}$$

As a result, the differential entropy $s[X]$ of the continuous random variable X is equal to the limit for $\Delta \to 0$ of the information entropy $S\left[X^{\Delta}\right]$ of its discretized counterpart X^{Δ} (*i.e.* the direct equivalent of $S[X]$ in Equation 1.35) up to an additive factor which is infinite since $\lim_{\Delta \to 0} \ln \Delta \to -\infty$. Therefore, differential entropy $s[X]$ is not well-defined as a proper extension of information entropy. Nonetheless, in practice, for a finite sample of X we compute the differential entropy choosing a fixed Δ as the bin width, *i.e.* without taking the limit $\Delta \to 0$. Similarly, the discretized information entropy for bins of variable width Δ_i is

$$\begin{aligned} S\left[X^{\mathbf{\Delta}}\right] &= -\sum_{i=-\infty}^{+\infty} p_i \ln p_i = -\sum_{i=-\infty}^{+\infty} f(x_i)\,\Delta_i \ln\left(f(x_i)\,\Delta_i\right) \\ &= -\sum_{i=-\infty}^{+\infty} f(x_i)\,\Delta_i \ln\left(f(x_i)\right) - \sum_{i=-\infty}^{+\infty} f(x_i)\,\Delta_i \ln(\Delta_i)\,. \end{aligned} \tag{1.42}$$

where the discretized version of X with variable bin widths $\mathbf{\Delta} = \{\Delta_i\}$ is

$$X^{\mathbf{\Delta}} = x_i \qquad \text{if} \quad i\Delta_i \leq X \leq (i+1)\Delta_i\,.$$

The probability p_i associated to the discretized variable $X^{\mathbf{\Delta}}$ in bin i is then

$$p_i = f(x_i)\,\Delta_i = \int_{i\Delta_i}^{(i+1)\Delta_i} f(x)\,dx. \tag{1.43}$$

The definition of the Kullback-Leibler divergence Equation 1.36 can also be extended to the probability density functions $f(x)$ and $g(x)$ as

$$d_{\mathrm{KL}}\left(f||g\right) = \mathbf{E}_f\left[\ln\left(\frac{f(x)}{g(x)}\right)\right] = \int_{-\infty}^{+\infty} f(x) \ln\left(\frac{f(x)}{g(x)}\right). \tag{1.44}$$

The same discretization detailed above can be applied to obtain the discretized version of the Kullback-Leibler divergence $d_{\mathrm{KL}}\left(f||g\right)$ from the probabilities p_i (see Equation 1.43) and q_i:

$$q_i = g(x_i)\,\Delta_i = \int_{i\Delta_i}^{(i+1)\Delta_i} g(x)\,dx.$$

which yields to

$$\begin{aligned} D_{\mathrm{KL}}\left(\mathbf{P}||\mathbf{Q}\right) &= \mathbf{E}_{\mathbf{P}}\left[\ln\left(\frac{\mathbf{P}\left(X^{\mathbf{\Delta}}\right)}{\mathbf{Q}\left(X^{\mathbf{\Delta}}\right)}\right)\right] \\ &= \sum_{i=-\infty}^{+\infty} p_i \ln\left(\frac{p_i}{q_i}\right) = \sum_{i=-\infty}^{+\infty} f(x_i)\,\Delta_i \ln\left(\frac{f(x_i)\,\Delta_i}{g(x_i)\,\Delta_i}\right) \\ &= \sum_{i=-\infty}^{+\infty} f(x_i)\,\Delta_i \ln\left(\frac{f(x_i)}{g(x_i)}\right). \end{aligned} \tag{1.45}$$

1.2.1 Maximum entropy principle

Information Theory provides a well-grounded conceptual framework to quantify the information content of a random variable through its entropy. However, in real-world situations, we may foresee that a random variable is not maximally haphazard, but has some peculiar features that determine typical expected values. In such case, the typical values control the functional form of the probability distribution followed by the random variable which, in general, is distinct from the uniform distribution. Indeed, the latter is the most unbiased distribution that satisfies only the normalization constraint over a finite interval. Likewise, we can impose other constraints that influence the functional form of the probability distribution of a random variable, given the expected values of quantities that depend on the variable itself. This is a very general problem in mathematical optimization that was first solved by Joseph Louis Lagrange in analytical mechanics [192]. The mathematical technique that he developed was named after him as the method of **Lagrange multipliers**. The statement of the problem is the following: we have a continuous function $f(x_1,\dots,x_n)$ of n variables $\{x_1,\dots,x_n\}$ which is subject to M equality constraints like $g_k(x_1,\dots,x_n)=0$ and we want to extremize it, finding the local maxima and minima.[2] To this aim, we introduce the so-called Lagrange multipliers λ_k [193–195], one for each constraint, that become the additional variables of an auxiliary function

$$\tilde{f}(x_1,\dots,x_n,\lambda_1,\dots,\lambda_M) = f(x_1,\dots,x_n) - \sum_{k=1}^{M} \lambda_k g_k(x_1,\dots,x_n)\,, \tag{1.46}$$

which incorporates the constrained functions $g_k(x_1,\dots,x_n)$. The method of Lagrange multipliers consist in solving the equations

$$\nabla_{x_1,\dots,x_n,\lambda_1,\dots,\lambda_M}\tilde{f}(x_1,\dots,x_n,\lambda_1,\dots,\lambda_M) = \nabla_{\boldsymbol{x},\boldsymbol{\lambda}}\tilde{f}(\boldsymbol{x},\boldsymbol{\lambda}) = 0\,, \tag{1.47}$$

where $\boldsymbol{x}=(x_1,\dots,x_n)$, $\boldsymbol{\lambda}=(\lambda_1,\dots,\lambda_M)$ and

$$\nabla_{x_1,\dots,x_n,\lambda_1,\dots,\lambda_M}\tilde{f} = \left(\frac{\partial\tilde{f}}{\partial x_1},\cdots,\frac{\partial\tilde{f}}{\partial x_n},\frac{\partial\tilde{f}}{\partial\lambda_1},\cdots,\frac{\partial\tilde{f}}{\partial\lambda_M}\right) = \left(\nabla_{\boldsymbol{x}}\tilde{f},\nabla_{\boldsymbol{\lambda}}\tilde{f}\right),$$

is the gradient of $\tilde{f}$ that includes the first partial derivatives of $\tilde{f}$ with respect to every variable. Equation 1.47 encapsulates $n+M$ equations in $n+M$ unknown variables, where each of the M terms $\frac{\partial}{\partial\lambda_k}\tilde{f}(x_1,\dots,x_n,\lambda_1,\dots,\lambda_M)$ results in $g_k(x_1,\dots,x_n)=0$, that is, the constraint associated to λ_k is satisfied. Hence, solving Equation 1.47 is equivalent to solve the system of equations

$$\begin{cases} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) - \sum_{k=1}^{M}\lambda_k\nabla_{\boldsymbol{x}} g_k(\boldsymbol{x}) = 0\,, \\ g_1(\boldsymbol{x}) = \cdots = g_M(\boldsymbol{x}) = 0\,. \end{cases} \tag{1.48}$$

where the first line is a shorthand for n equations and the second line represents the M equations of the constraints. It is worth to remark that the fixed constraints g_k uniquely determine the

[2] The formulation of the extremization problem is well-defined only if we assume that the function f and the constraints g_k, as well as their first derivatives, are continuous.

functional form of f after the system in Equation 1.48 is solved.

Here, we use the method of Lagrange multipliers to determine the probability distribution of a random variable that is consistent with known constraints but is maximally unbiased otherwise. As we motivated at the beginning of the Section, information entropy is the proper measure to quantify the uncertainty or ignorance about the distribution of a random variable: if the probability is more biased toward a value, the entropy of the random variable is lower; on the contrary, the entropy is maximal if the random variable is uniformly distributed. Entropy, which depends on the probability distribution, is then the right function to maximize in order to establish the most even probability distribution that satisfies the expected values as constraints. The maximum entropy principle was first proposed in the milestone paper [196] of Edwin T. Jaynes as a mindful approach to recover the rules of statistical mechanics in terms of an inference problem from partial knowledge. In this way, the maximum entropy principle allows the derivation of the maximum entropy probability distribution, which is the best possible estimate based on a limited knowledge represented by the constraints, being at the same time the least biased estimate. Accordingly, the maximum entropy distribution is the most ignorant choice that could be made respecting only the constraints, as we do not construct it from knowledge that we do not have.

As a simple application of the maximum entropy principle, consider the discrete random variable X whose possible outcomes are $\{x_1, x_2, \dots, x_n\}$ with associated probabilities $\boldsymbol{p} = (p_1, \dots, p_n)$ where $p_i = P(X = x_i)$ for every $i = 1, \dots, n$. The least constrained probability mass function is the one that maximize the information entropy in Equation 1.35: following the strategy of the Lagrange multipliers, the function f in Equation 1.46 is the information entropy, which depends on the probabilities p_i:

$$f(p_1, p_2, \dots, p_n) = -\sum_{i=1}^{n} p_i \ln p_i \,. \tag{1.49}$$

If no expected values are imposed as actual constraints, the only rule that the probability mass function must obey is the normalization condition $\sum_{i=1}^{n} p_i = 1$, which can be regarded as the only constraint to fulfill:

$$g(p_1, p_2, \dots, p_n) = \sum_{i=1}^{n} p_i - 1 \,. \tag{1.50}$$

The system in Equation 1.48 then becomes

$$\begin{cases} \nabla_{\boldsymbol{p}} \left(-\sum_{i=1}^{n} p_i \ln p_i \right) - \nu \nabla_{\boldsymbol{p}} \left(\sum_{i=1}^{n} p_i - 1 \right) = 0 \,, \\ \frac{\partial}{\partial \nu} \tilde{f}(\boldsymbol{p}, \nu) = g(\boldsymbol{p}) = \sum_{i=1}^{n} p_i - 1 = 0 \,. \end{cases} \tag{1.51}$$

In this example, ν is the Lagrange multiplier associated to the constraint on the normalization of the probability mass function. Thus, every single equation among n obeys

$$\frac{\partial}{\partial p_k} \left\{ -\sum_{i=1}^{n} p_i \ln p_i - \nu \left(\sum_{i=1}^{n} p_i - 1 \right) \right\} = -\ln p_k - 1 - \nu = 0 \,. \tag{1.52}$$

The probability mass function is then

$$p_k = e^{-\nu-1}, \tag{1.53}$$

by imposing the normalization constraint of Equation 1.50, we get

$$\sum_{k=1}^{n} p_k = \sum_{k=1}^{n} e^{-\nu-1} = ne^{-\nu-1} = 1,$$

therefore

$$p_k = e^{-\nu-1} = \frac{1}{n}, \tag{1.54}$$

The least constrained maximum entropy probability distribution is the uniform one.

We detailed here the motivations to adopt the maximum entropy principle when considering probability distributions with peculiar traits since it will be used in the next Chapter to characterize the frequency of occurrence of words within texts. In such case, the probability distribution tends to have a broad profile, similar to a power-law, an ubiquitous distribution that describes a plethora of critical phenomena in nature, ranging from the income of people to the intensity of earthquakes [197].

2 Entropic selection of concepts in networks of similarity between articles

2.1 Analysis of scientific articles

In this Chapter we will make use of the concepts introduced in the previous one to study the relationships amid scientific manuscripts. Since articles are the main sources of dissemination of scientific knowledge, describing their interdependencies is of primary importance to understand the fundamental principles that shape science. The more natural approach to investigate their relationships it to consider articles as basic elements that constitute the nodes in a network. The most recurrent policy to create links between articles is from the reference list, in such a way that an article is connected to all the articles that are cited within. Given the immediate definition of the links, the resulting *citation network* has been among the first attempts to quantitatively sketch a map of scientific knowledge in the seminal work of de Solla Price [15]. Thanks to the availability of machine-readable bibliometric data about the articles, the studies based on such resources have been flourishing addressing various aspects of the citation patterns, from the simple empirical characterization of the citation distribution of papers [53], the discovery of its universal features [198] and their utility to gauge the impact of the publications [199] to the modeling of the citation mechanisms [200], the spreading of recurrent memes in the scientific literature [4] and effect of the temporal dimension of citations in different fields [201]. Finally, bibliometric data have been adopted to construct several maps of the scientific landscape in order to conduct exploratory data analyses of the their information [61, 62, 64, 202].

By construction, citation networks clearly acknowledge important contributions from papers in the existing scientific literature [203]. However, the ever-increasing number of papers that are published yearly represent a major issue for scientists that need to keep up-to-date with recent advances in their fields [23–25]. This problem is even more pronounced for interdisciplinary research where the number of venues to watch over easily expands. Concurrently, the chance of missing relevant work due to the scattering of papers in several venues likely grows [204]. A paradigmatic example is provided by the field of complex networks [205], a truly multidisciplinary community where contributions spread across scholars with very different backgrounds and approaches, ranging from physicists to sociologists and economists. Therefore, thinking of

reading all the new manuscripts that are published is not feasible since it would take too much time and efforts. Ideally, the optimal solution would be to focus only on those papers that are relevant. In order to help scholars in such activity, various tools have been designed throughout the years [206]. A primary source to suggest interesting articles is following the citation paths from other articles. Although citations remain a trustful source to point out previous knowledge, their potential is limited to the subjectivity of the authors. Therefore, a more effective approach to highlight similar studies relies on the semantic analysis of articles which may potentially uncover related topics within. Usually, a large-scale analysis is performed from similarities between titles and abstracts of the articles [66, 207] with the aid of natural language processing techniques that allows to automatically extract keywords from these elements [208]. On the one hand, the careful choice of the words adopted within eventually capture precise similarity patterns between articles. On the other hand, the concise nature of title and abstracts likely overlook deeper terms nested within texts that potentially contribute to the similarities with other articles. For that reason, a thoughtful procedure is to exploit the full text of articles (including the body) to semantically relate articles one another [65–67, 209, 210]. Of course, this approach is not perfect and possesses its own limitations; for example, it cannot overcome the differences between jargons that characterize the same concepts which are often field-specific [211]. Nevertheless, the semantic analysis of the article texts may potentially uncover similarities between topics in different articles. From the semantic similarity network, the thematic organization of manuscripts emerges as groups of articles related to similar topics. Exploring such organization is then useful to map the actual structure of the scientific knowledge [10, 60, 212].

Unfortunately, current methodologies to automatically extract keywords do not consider the specificity of scientific terminology. Thanks to the collaboration with the ScienceWISE (SW) team, we had access to a crowdsourced ontology of *scientific concepts* available in the SW platform[1]. Such ontology has been extracted for a set of scientific manuscripts appearing in the `arXiv` [2] electronic preprint repository. Indeed, the aim of the SW project is to help scientists organizing their personal collection of preprints, enriching it through semantic tagging and recommending interesting articles [213–218].

The manuscripts on `arXiv` are organized in categories that span several domains, from physics to economics. The categories, however, are not static but evolve in time, continuously adapting according to the feedback of the community of researchers that use the repository. As an example, in its prime astrophysics was a unique category but after some time got split into six subcategories[3] with the purpose of differentiating the submissions into finer-grained subjects. In the same spirit of adaptation, new categories have been introduced over the twenty-year history of the repository, *e.g.* quantitative biology and finance, in order to encourage researchers to disseminate their work [25]. During the submission process of a manuscript, the authors must assign it to a primary (sub)category with the optional choice of cross-listing to other secondary categories. However, the different (sub)categories are not pretended to be rigorous in providing a

[1] http://sciencewise.info

[2] https://arxiv.org/

[3] In the `arXiv` nomenclature, the subcategories are referred to as subject classes

principled division of a field in several areas with the same specificity and importance. In the field of physics, for example, high energy physics is divided in four subcategories with no further subdivisions, while the physics category includes very disparate subcategories, from medical physics to atmospheric and ocean physics.

Among all the possible choices available, we selected articles that have been submitted during the year 2013 under one of the physics categories as primary subject, independently on their secondary ones. The resulting corpus consists of 52979 articles, whose composition in terms of `arXiv` categories is reported in Table 2.1 and Figure 2.1.

Table 2.1 – Number of articles N_a and relative size (in %) of each `arXiv` category in the physics corpus. The first and second column include the name and the abbreviation of the category. If a given category is not the standard one but results from merging multiple categories, their suffix is indicated within brackets after the abbreviation.

Category	Abbreviation	N_a	%
Condensed matter	cond-mat	12679	23.93
Astrophysics	astro-ph	12458	23.51
High energy physics	hep [-ex, -lat, -ph, -th]	9661	18.24
Physics	physics	7407	13.98
Quantum physics	quant-ph	4039	7.62
General relativity and quantum cosmology	gr-qc	2273	4.29
Nuclear physics	nucl [-ex, -th]	1819	3.43
Mathematical physics	math-ph	1767	3.34
Nonlinear sciences	nlin	876	1.65
	Total	52979	100

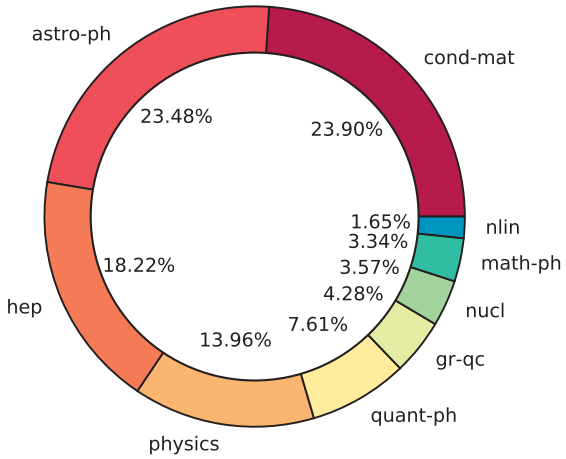

Figure 2.1 – Donut chart of the composition of the physics articles in terms of `arXiv` primary categories.

The donut chart in Figure 2.1 shows that the corpus is highly heterogeneous in terms of the number of articles. Indeed, the two most populated categories, cond-mat and astro-ph, comprise

together almost half of the total number of articles, while gr-qc, nucl, math-ph and nlin do not even size up to 13%. In order to distillate the content of the articles, we mine the texts as follows: first, as a preliminary step, we preprocess the texts to get rid of the so-called "stop words", namely terms that are syntactic elements present in any written text and do not carry any semantic meaning. Examples of such terms are articles (*e.g.* , "the", "this" and "any"), conjunctions (*e.g.* , "and", "but" and "after") and adverbs (*e.g.* , "usually", "very" and "enough"). Next, we identify important words using the keyword extraction algorithm KPEX [219]. Finally, we match the keywords with an ontology of scientific concepts that is accessible on the SW platform. These concepts have been gathered initially from online encyclopedias and subsequently curated by crowdsourcing from SW users expert in different areas of physics. Note that the availability of a curated ontology for a large dataset is not common and can be regarded as a special feature of our dataset. Overall, the articles in physics contain 11,637 unique concepts: from this pool, we removed concepts that are present only in one article (since they do not contribute to the similarity between articles), and those occurring always the same number of times inside the articles, eventually obtaining 10,661 concepts. Among them, 339 have been marked as *common* by experts. To characterize the relationships between articles in terms of the words that they share, we then analyze the unipartite projection onto the articles as detailed in subsection 1.1.2, capturing with the link weights the similarities between documents. The only relationships that are present between them are described by the link weights that quantify the relevance of a concept c in a document α. Several policies can be adopted to define such relevance. We decided to adopt the *de facto* standard measure in information retrieval is the TF-IDF [220, 221]. The TF-IDF of a concept c in article α is defined as the product of two quantities: the *boosted term-frequency*, $tf_c(\alpha)$, which is the number of occurrences of concept c in article α magnified depending on its position in the text of the article, and the *inverse document frequency*, IDF_c, which corresponds to the inverse fraction of articles where concept c appears. The rationale behind the TF-IDF is indeed the following: on the one hand, the *boosted term-frequency* is a local measure, at the level of single articles, which gives importance to prevalence of a concept inside a given text; on the other hand, the *inverse document frequency* is a global measure, at the level of the corpus of articles, which penalizes a concept appearing in many articles. In order to compute the *boosted term-frequency*, the full text of every article α is split in 3 different parts: the title t, the abstract a and the body b. For each part $p = \{t, a, b\}$ we compute the number of times the concept c is found in it, $n_c^p(\alpha)$. The *boosted term-frequency* is then calculated as

$$tf_c(\alpha) = 5 \cdot n_c^t(\alpha) + 3 \cdot n_c^a(\alpha) + n_c^b(\alpha)\,. \tag{2.1}$$

The *inverse document frequency* of concept c is

$$IDF_c = \ln\left(\frac{N_a}{N_c}\right), \tag{2.2}$$

where N_a is the total number of articles in the corpus and N_c is the number of articles that contain concept c. The content of the article α is then summarized by a vector $\vec{d}_\alpha$ whose elements

correspond to the TF-IDF of the concepts in the article α:

$$d_\alpha(c) = \begin{cases} tf_c(\alpha) \cdot IDF_c & \text{if } c \in \mathcal{C}_\alpha\,, \\ 0 & \text{otherwise,} \end{cases} \tag{2.3}$$

where $\mathcal{C}_\alpha$ is the set of concepts contained in article α. The similarity between a pair of articles α and β is represented by the link weight in the unipartite projection onto the articles, as defined in Equation 1.1, by using $\vec{d}_\alpha$ and $\vec{d}_\beta$:

$$w^{\mathcal{P}}_{\alpha\beta} = \frac{\vec{d}_\alpha \cdot \vec{d}_\beta}{\left\|\vec{d}_\alpha\right\| \left\|\vec{d}_\beta\right\|}\,, \tag{2.4}$$

where $\cdot$ denotes the scalar product and $\|\ldots\|$ is the Euclidean norm. Equation 2.4 is the definition of *cosine similarity* between vectors, which is equivalent to imposing $\boldsymbol{a}_{i:} = \frac{\vec{d}_\alpha}{\left\|\vec{d}_\alpha\right\|}$ and $\boldsymbol{a}_{j:} = \frac{\vec{d}_\beta}{\left\|\vec{d}_\beta\right\|}$ in Equation 1.1. The cosine similarity between normalized vectors has the advantage that the link weight $w^{\mathcal{P}}_{\alpha\beta} \in [0, 1]$, where $w^{\mathcal{P}}_{\alpha\beta} = 0$ indicates that the articles are not sharing any concept at all (*i.e.* are completely different) since the vectors form an angle $\theta = 90°$in the space of the concepts. On the contrary, a value $w^{\mathcal{P}}_{\alpha\beta} = 1$ is found if the documents do not simply share the same set of concepts, but use them with the equal occurrences up to a multiplicative constant (*i.e.* they are identical); therefore, the corresponding vectors form an angle $\theta = 0°$. As links with very low weight are likely to include spurious similarities, we discard from the network all the links whose weight is immaterial, *i.e.* $w^{\mathcal{P}}_{\alpha\beta} \leq w_{min}$, fixing a threshold $w_{min} = 0.01$ which corresponds to an angle $\theta_{min} = 89.43°$. A general description of the similarities between articles is achieved from the structural analysis of the network using the topological quantities introduced in subsection 1.1.3. An important measure to gauge the overall connectivity is the average degree $\langle k \rangle = 19334$: this value is of the same order of the size of the corpus $N_a = 52979$, therefore it is extremely high if compared to other real-world networks [81]. Likewise, the sparsity of the network measured by the link density is $\rho = 36.5\%$, indicating a very dense pattern of connections. Even if these quantities are best suited to characterize a (simple) unweighted network, they make clear the presence of an overwhelming number of links. This circumstance undermines the modeling of the system as a network since one of the main benefits of this approach is to deal with sparse interactions. In the next Section, we propose a solution to the problem of limiting the number of connections of the similarity network. The majority of the work presented in this Chapter and the related Appendix has been the subject of [222, 223].

2.2 Sparsifying the similarities between articles

The necessity of pruning the links of a weighted network is a common requirement in many real-world cases. Several techniques have been suggested to tackle this problem by preserving the significant links of a weighted network, as described in subsection 1.1.6. However, they operate *a posteriori* on the weight distribution without considering the specificity of the relations encoded

in the weights. Indeed, the projection of the bipartite network onto the articles $\mathcal{P}$ entails a loss of information. Instead of filtering the weights, a more well-grounded approach consists in retaining only *relevant* concepts before computing the similarities, thus acting *ex-ante* on the procedure that generates the weights. Loosely speaking, the concepts that should be discarded are those that are pervasive but do not carry a specific meaning, nor are they related to a particular domain. These concepts are the so-called "*common concepts*" (CCs) which inflate the link weights between articles and are responsible for the spurious similarities.

As explained in the previous Section, a built-in set of CCs is already present in the SW platform. Users have been asked to tag the crowdsourced concepts that they believe are common or suggest new ones as such. However, maintaining an updated set of CCs requires a periodical collaboration from the experts. Indeed, the SW platform is directly linked to `arXiv` as it includes new articles on a daily basis. The examination of their concepts can then become quite demanding for the experts. Moreover, the identification of CCs is based only on the judgment of the experts that are unaware of the composition of the whole corpus in terms of subjects. This point is crucial because a concept like *graphene* could be regarded as common for a corpus of articles on material science but it is likely specific for a corpus on astrophysics. These drawbacks suggest that removing only the CCs may not be a thoughtful choice. Nevertheless, we can leverage such information to pinpoint common traits of CCs, and use such features identify "hidden" common concepts that are not been tagged as such. Concepts that are not considered as common are then relevant and will be used to construct a "purified" version of the similarity network between articles.

The question that arises is then the following: is there a method that allows to automatically spot relevant concepts which depends, by construction, on the corpus under scrutiny? In order to give an answer, we first need to define the fingerprints of a *relevant* concept. A concept of this kind should satisfy two requirements, at least:

1. It must neither be too widespread (*i.e.* a buzzword) nor too rare (too specific) among the articles in the corpus.

2. It must occur a considerable number of times within articles.

The first feature describes the *discriminative* power of a concept as it appears in a significant number of papers but it is also useful to discern articles. The second feature accounts for the *pertinence* of a concept in delineating the content of an article. For a corpus of N_a articles, the number of unique concepts is defined as the union of the concepts that are present in the different articles, $\mathcal{C} = \bigcup_{\alpha=1}^{N_a} \mathcal{C}_\alpha$. The discriminative power of a concept $c \in \mathcal{C}$ that is present in N_c articles is measured by its *document frequency*[4] $df_c = \frac{N_c}{N_a}$, while the pertinence for a given article α is calculated as the (boosted) *term-frequency* $tf_c(\alpha)$. The average term-frequency of c then reads

[4] In the following, we will use the same document frequency df_c as a symbol to denote the raw number of articles N_c where concept c appears in. A distinct definition is not considered since the two quantities are the same up to a divisive constant, thus they have the same properties and can be interchanged without affecting the conclusions drawn, *e.g.* , for the their distribution.

$\langle tf_c \rangle = \frac{1}{N_c} \sum_{\alpha=1}^{N_c} tf_c(\alpha)$. The characterization of the concepts is pursued in terms of their df and $\langle tf \rangle$ by partitioning the corresponding plane according to several thresholds on these quantities. The classification of the concepts is then established from the regions that they occupy, defined as follows:

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

A2 The domain of *common/ubiquitous* concepts showing high values of both df and $\langle tf \rangle$.

A3 The domain of *relevant/informative* concepts that have intermediate values of df and $\langle tf \rangle$.

A4 All the residual concepts that seldom appear within documents (on average) to be regarded as relevant.

Considering each trait individually, we can naively split its range in three, or more, intervals that include low, medium and high values. The boundaries of the domains are then defined in terms of the percentiles in order to capture at best the variability of the data. In particular, we consider three intervals for $\langle tf \rangle$ delimited by 25th and 75th percentiles, and four intervals for df delimited by the 25th, 75th and 90th percentiles. The resulting partitioning is displayed in Figure 2.2. As

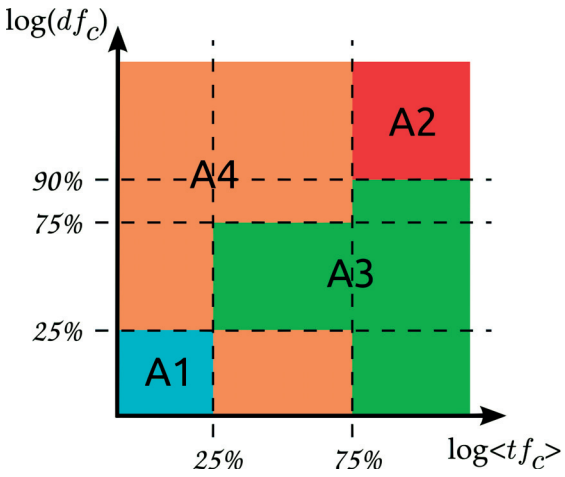

Figure 2.2 – Classification of the concepts based on the tessellation of the $(\langle tf \rangle, df)$ plane. Each type of concept is identified by a different color. The dashed lines correspond to the different percentiles that delimit the domains.

can be appreciated in Figure 2.3, the actual position of the concepts in the plane indicate that both variables span a wide range of values. The type of concepts is denoted by the color: green dots, for example, identify the concepts defined as relevant by the partitioning scheme, which amount to 46% of the total. In principle, we can keep only these concepts to calculate the similarity between articles. Nevertheless, the intuitive tessellation of the plane has several drawbacks: indeed, it depends on many thresholds that are arbitrary both in their number and value. Furthermore, concepts hand-marked as common by experts (the black points in Figure 2.3) are not located in

any specific region of the $(\langle tf \rangle, df)$ plane. Still, they tend to concentrate in a band that is not a straightforward combination of df and/or $\langle tf \rangle$, as shown by the black curve. Curiously, this trend already suggests the presence of a common principle that influence their position. Last but

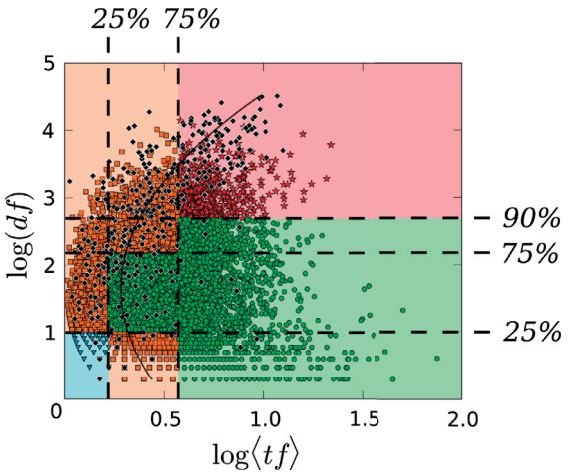

Figure 2.3 – Bidimensional tessellation of the concepts represented by different shapes and colors according to the domain they are assigned to, as reported in Figure 2.2. Ubiquitous (red stars), rare/specific (cyan triangles), significant (green circles), and other concepts (orange squares) constitutes the 4.6, 11, 46, and 39 % of the total, respectively. Black points denote concepts tagged as common (CCs) and the solid black line is a guide for the eye that interpolate their average position. Logarithmic scales are adopted to neatly visualize the results.

not least, both quantities have a broad distribution as displayed in Figure 2.4. This property is not surprising since has been already observed in [224–228]. Besides, it is intimately related to a similar behavior observed for the total number of word repetitions in a corpus which obeys the Zipf's law [229]. Scale-free quantities do not possess a characteristic scale, thus imposing some threshold is not only subjective but also inappropriate. These shortcomings suggest the pursuit of an alternative method to filter concepts which should not rely on simple quantities associated to them but, ideally, on some *microscopic characteristic* of the concepts themselves. More precisely, every concept has a specific distribution of the term-frequency over the articles that describes how it is used. Therefore, we are interested to understand more in detail the peculiar traits of such distributions in order to characterize the concepts. To this aim, we can regard the term-frequency of a concept as a random variable. A suitable measure to quantify its information content is given by the entropy introduced in section 1.2. Thus, every concept has an associated entropy, S_c, that can be used as a proxy for its importance [230–232]:

$$S_c = -\sum_{k=1}^{\infty} p_c(k) \ln p_c(k), \tag{2.5}$$

where $p_c(k) = \frac{N_c(k)}{N_c}$ is the ratio between the number of articles where concept c occurs k times, $N_c(k)$, and the total number of articles it appears in, N_c. In other words, $p_c(k)$ is the probability that a document picked at random among N_c contains k occurrences of concept c. The entropy

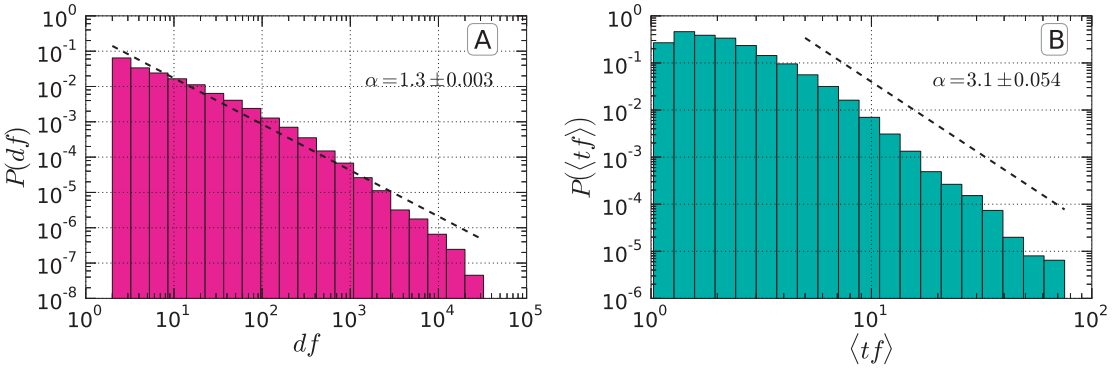

Figure 2.4 – Probability density function of the concept features: panel (A) refers to the document frequency, df, while panel (B) is related to the average term-frequency, $\langle tf \rangle$. In each panel, the power-law fit of the corresponding distribution, $P(x) \sim x^{-\alpha}$, is displayed by a dashed line along with the parameter α and its standard deviation. The fits are reported only to highlight the broad shape of the distributions and they are not intended to represent the best fitting models. The trend of the distributions reveal the great variability of such quantities that may lack a typical scale.

S_c is called *conditional* since it is calculated under the condition that the concept c is present within the articles. We then investigate the relation between S_c and the emblematic attributes of a concept, (*i.e.* df_c and $\langle tf_c \rangle$). Intuitively, we may expect that concepts present in many articles are likely to be classified as common. However, there is no evidence that confirms this claim. Indeed, Figure 2.5 A illustrates the df of concepts are present as a function of the entropy, S_c. CCs, denoted by black points, do not exhibit any shift toward particular values of S_c, but are evenly scattered across the whole range (as already noted in Figure 2.3). Nonetheless, a linear correlation between $\log_{10}(df)$ and S_c exists, as shown by the black line, implying that the two quantities are somehow related. However, if we consider the $\langle tf \rangle$, CCs tend to have an higher S_c when compared to the other concepts with the same $\langle tf_c \rangle$, as displayed in Figure 2.5 B. Furthermore, CCs are concentrated toward an ideal border of the concept distribution in the $(S_c, \langle tf \rangle)$ plane. The existence of such limit region is quantitatively demonstrated by the dashed line describing the maximum entropy computed after imposing $\langle tf \rangle$ as a constraint:

$$S_{\langle tf \rangle} = \langle tf \rangle \ln\big(\langle tf \rangle\big) - \big(\langle tf \rangle - 1\big) \ln\big(\langle tf \rangle - 1\big)\,. \tag{2.6}$$

For brevity, the derivation of such formula is given in subsubsection A.1.2.1. Likewise, a condensation of S_c for CCs is noticed also in the case of $\langle \ln\big(tf\big) \rangle$ (Figure 2.5 C) where such trend is even more pronounced with respect to the one observed for $\langle tf \rangle$. Therefore, the peculiar behavior of S_c as a function of $\langle tf \rangle$ and $\langle \ln\big(tf\big) \rangle$ suggests that they are both intimately related to the high entropy of CCs. However, the raw value of S_c is not sufficient to discriminate CCs. Indeed, we need to fairly assess if the observed S_c is big or not with respect to a reference value. Given the trends recognized above, we argue that CCs are driven close to their maximum possible entropy because of some fundamental mechanism. Then, a proper comparison of the actual entropy should be made considering the expected maximum entropy as its theoretical counterpart. The adoption of the *maximum entropy principle*, however, is not only motivated by empirical

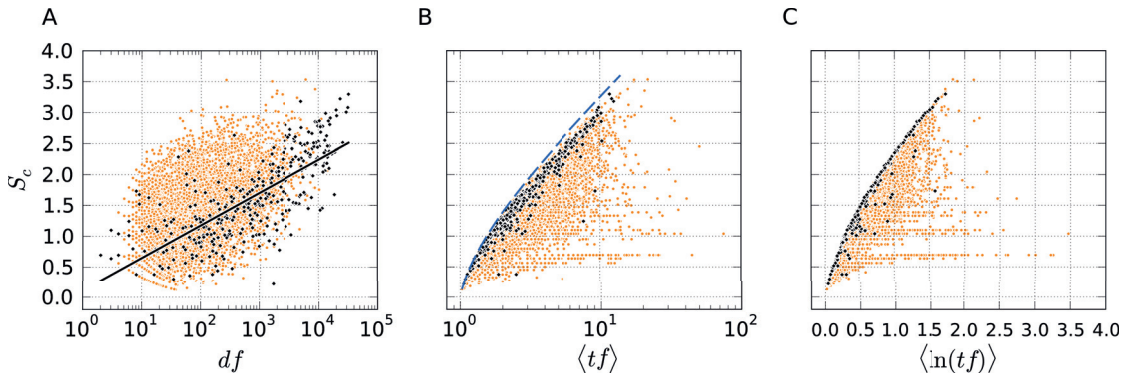

Figure 2.5 – Relations between the entropy S_c and several features of the concepts. Common concepts in SW (CCs) are denoted by black points. (A) Dependence between S_c and the number of articles where concepts appear in, df. The solid line is the linear least-squares regression between $\log_{10} df$ and S_c for CCs (Pearson correlation coefficient $r = 0.743$). (B) Dependence between S_c and the average term-frequency, $\langle tf \rangle$. The dashed line represents the analytical expression of the maximum entropy as a function of $\langle tf \rangle$, calculated by imposing only the constraint on $\langle tf \rangle$ itself. (C) Dependence between S_c and the average logarithm of the term-frequency, $\langle \ln(tf) \rangle$.

evidence but it is also justified from theoretical arguments, as we detailed in subsection 1.2.1.

In the scientific literature, an increasing interest has been devoted to the investigation of the statistical properties of word [224–227]. Specifically, the distribution of the term-frequency of a word is described by a law spanning a broad range of values. In order to establish the constraints that are more appropriate to recreate the word frequency distribution, a useful hint is given by the empirical findings unveiled above. The average term-frequency, $\langle tf \rangle$, and the average of the logarithm of the term-frequency, $\langle \ln(tf) \rangle$, may be suitable constraints as they are clearly related to the high entropy S_c of common concepts[5]. The inspection of the term-frequency distribution of various concepts (in particular the common ones), as displayed in Figure 2.6, reveals that $p_c(k)$ is well characterized by a power-law with a cutoff

$$q_c(k) \propto k^{-s} e^{-\lambda k}. \tag{2.7}$$

represented by the dashed line in Figure 2.6. Equation 2.7 is precisely the functional form of the maximum entropy probability distribution that follows from such constraints [226]. Furthermore, it has been already adopted to model the tf distribution of words [225]. The maximum entropy principle described in subsection 1.2.1 allows to derive the analytical expression for the expected distribution $q_c(k)$ as in Equation 2.7. In particular, the maximization of the entropy S is performed under the constraints that the first moment and log-moment of the term-frequency k must match

[5] Naively, the choice of the constraints can be explained in the following way: first, the mean value of a variable is its most characteristic feature, therefore is reasonable to select $\langle tf_c \rangle$ as a constraint of the maximum entropy distribution. Second, many studies confirm that the term-frequency distribution is spread across a wide range of values, exhibiting a heavy tail profile [225, 227]. The proper constraint to reproduce such behavior is then the average of the logarithm of the variable, $\langle \ln(tf_c) \rangle$, which can be considered as the simplest typical value on a broader (logarithmic) scale.

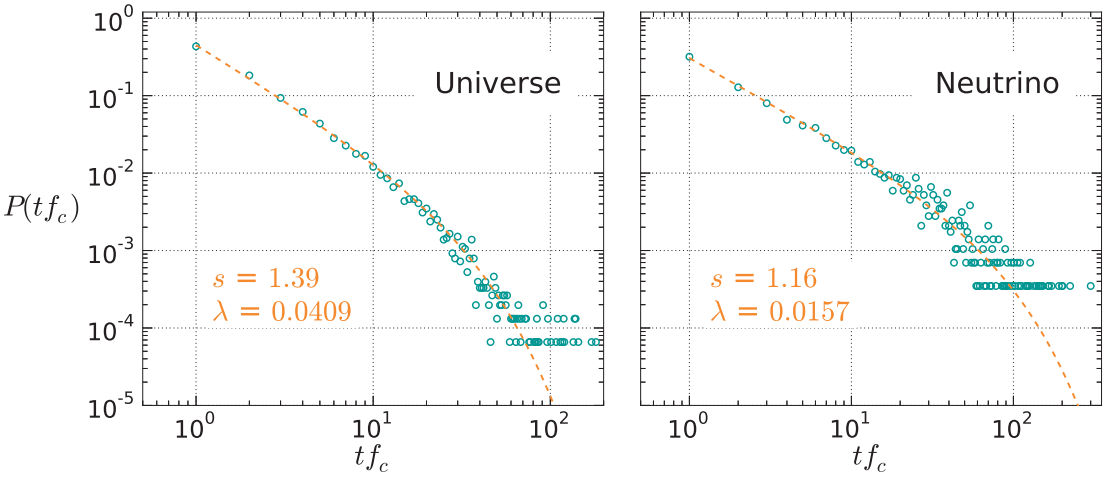

Figure 2.6 – Typical distributions of the term-frequency, tf_c, for two representative concepts. For each concept, the dashed line indicates the maximum entropy distribution, a power-law with a cutoff, whose parameters s and λ are reported. (A) Concept "Universe" is an example of a common concept. (B) Concept "Neutrino" is taken from those concepts identified as generic by the entropic filtering, as explained later.

the empirical values, $\langle tf_c \rangle$ and $\langle \ln(tf_c) \rangle$ respectively. As a consequence, the expression to maximize reads

$$S_{max} = -\sum_{k=1}^{\infty} q_c(k) \ln q_c(k) - \lambda \left(\sum_{k=1}^{\infty} k\, q_c(k) - \langle tf_c \rangle \right) - s \left(\sum_{k=1}^{\infty} \ln(k)\, q_c(k) - \langle \ln(tf_c) \rangle \right) - \nu \left(\sum_{k=1}^{\infty} q_c(k) - 1 \right). \tag{2.8}$$

In this equation, λ is the Lagrange multiplier associated to the $\langle tf_c \rangle$ constraint, s is the one associated to $\langle \ln(tf_c) \rangle$ and ν is associated to the normalization condition of the probability mass function $q_c(k)$. The maximization of Equation 2.8 with respect to $q_c(k)$ is performed as $\frac{\partial S_{max}}{\partial q_c(k)} = 0$, which gives:

$$-\ln q_c(k) - 1 - \lambda k - s \ln(k) - \nu = 0\,. \tag{2.9}$$

Thus, the probability mass function q_c is defined as

$$q_c(k) = \frac{e^{-(\nu+1)} e^{-\lambda k}}{k^s}\,. \tag{2.10}$$

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 articles under scrutiny. Maximizing Equation 2.8 with respect to each Lagrangian multiplier allows to impose the relative constraint. As a consequence, the parameters that appear in Equation 2.10 are computed from the constraints. The maximization of Equation 2.8 with

respect to ν, $\frac{\partial S_{max}}{\partial \nu} = 0$, allows to recover the normalization condition:

$$\sum_{k=1}^{\infty} q_c(k) = e^{-(\nu+1)} \sum_{k=1}^{\infty} \frac{e^{-\lambda k}}{k^s} = 1\,,$$
$$e^{(\nu+1)} = \sum_{k=1}^{\infty} \frac{e^{-\lambda k}}{k^s} = \mathrm{Li}_s(e^{-\lambda})\,. \tag{2.11}$$

In the last equation, the infinite summation is equal to the special function called polylogarithm[6] of order s and argument $e^{-\lambda}$. Equation 2.11 allows to properly normalize the probability mass function in Equation 2.10 so that we obtain

$$q_c(k; s, \lambda) = \frac{\frac{e^{-\lambda k}}{k^s}}{\mathrm{Li}_s(e^{-\lambda})}\,, \tag{2.12}$$

the same expression reported in Equation 2.7. Maximizing Equation 2.8 with respect to λ, $\frac{\partial S_{max}}{\partial \lambda} = 0$, we recover the constraint $\langle t f_c \rangle$:

$$\sum_{k=1}^{\infty} k\, q_c(k) = \frac{\sum_{k=1}^{\infty} \frac{k\, e^{-\lambda k}}{k^s}}{\mathrm{Li}_s(e^{-\lambda})} = \frac{\sum_{k=1}^{\infty} \frac{e^{-\lambda k}}{k^{s-1}}}{\mathrm{Li}_s(e^{-\lambda})} = \langle t f_c \rangle\,,$$
$$\frac{\mathrm{Li}_{s-1}(e^{-\lambda})}{\mathrm{Li}_s(e^{-\lambda})} = \langle t f_c \rangle\,, \tag{2.13}$$

where in the last equation we applied the definition of the polylogarithm in Equation 2.11. Finally, the maximization of Equation 2.8 with respect to s, $\frac{\partial S_{max}}{\partial s} = 0$, allows to impose the constraint on $\langle \ln(t f_c) \rangle$:

$$\sum_{k=1}^{\infty} \ln(k)\, q_c(k) = \frac{\sum_{k=1}^{\infty} \frac{\ln(k)\, e^{-\lambda k}}{k^s}}{\mathrm{Li}_s(e^{-\lambda})} = \langle \ln(t f_c) \rangle\,,$$
$$-\frac{\partial_s \mathrm{Li}_s(e^{-\lambda})}{\mathrm{Li}_s(e^{-\lambda})} = \langle \ln(t f_c) \rangle\,. \tag{2.14}$$

The expression in the last equation has been derived thanks to the identity

$$\sum_{k=1}^{\infty} \frac{\ln(k)\, e^{-\lambda k}}{k^s} = -\frac{\partial}{\partial s} \sum_{k=1}^{\infty} \frac{e^{-\lambda k}}{k^s} = -\frac{\partial}{\partial s} \mathrm{Li}_s(e^{-\lambda})\,.$$

Considering that the two constraints in Eqs. (2.13) and (2.14) must be valid simultaneously, the

[6] For any value of $s, e^{-\lambda} \in \mathbb{C}$, the definition of the infinite sum as polylogarithm is limited to the case when the modulus of the argument is smaller than one, $|e^{-\lambda}| < 1$. Note, however, that in the present case only real valued parameters are meaningful.

resulting system of equations to solve is then

$$\begin{aligned} \frac{\mathrm{Li}_{s-1}(e^{-\lambda})}{\mathrm{Li}_s(e^{-\lambda})} &= \langle tf_c \rangle , \\ -\frac{\partial_s \mathrm{Li}_s(e^{-\lambda})}{\mathrm{Li}_s(e^{-\lambda})} &= \langle \ln\left(tf_c\right) \rangle . \end{aligned} \tag{2.15}$$

In Equation 2.15, both parameters s and λ are present in each of them. Since the two equations are coupled, the parameters cannot be calculated explicitly but we resort to solve the system numerically. The details of the algorithmic implementation of the system, along with some snippets of code, are provided in subsection A.3.1.

The analysis of the distribution of s and λ reveals the trend followed by the maximum entropy distribution q_c of the concepts, as illustrated in Figure 2.7. Interestingly, panel A highlights that $s = 3/2$ is the most recurrent value of the power-law exponent, therefore it is somehow characteristic of the mechanism that steers the empirical distributions of the term-frequency. Moreover, panel B shows the distribution of the exponential cutoff λ which is peaked around 0, highlighting that the cutoff mildly affects the power-law behavior in general. For every concept,

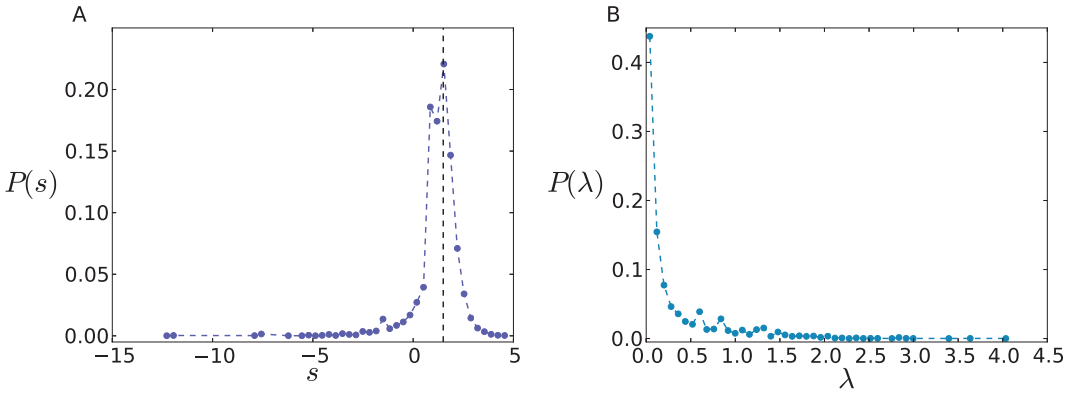

Figure 2.7 – Histogram of the parameters of the maximum entropy distribution, a power-law with a cutoff (see Equation 2.7), for the term-frequency of the concepts. (A) Distribution of the power-law exponent s that shows a pronounced peak at $s = 3/2$, as indicated by the dashed vertical line. Note that most of the values tend to be located around such maximum. (B) Distribution of the exponential cutoff λ which is squeezed toward zero.

the maximum entropy S_{max} associated to the probability mass function in Equation 2.12 is then

$$\begin{aligned} S_{max} &= -\sum_{k=1}^{\infty} q_c(k) \ln q_c(k) \\ &= \ln\left[\mathrm{Li}_s(e^{-\lambda})\right] + \lambda \langle tf_c \rangle + s \langle \ln\left(tf_c\right) \rangle . \end{aligned} \tag{2.16}$$

This expression is then the sum of three contributions, each coming from a Lagrange multiplier

which modulates the respective constraint. The interplay between the maximum entropy, S_{max}, and the observed entropy, S_c, is then examined in Figure 2.8 where we observe a clear correlation between such quantities. Indeed, the majority of common concepts (represented in black) tend to concentrate in a narrow region close to the line $S_c = S_{max}$. As a consequence, CCs displays a very high correlation between S_{max} and S_c. Comparing this trend with the relation observed for S_c against df (Figure 2.5 A), we immediately appreciate that S_{max} is more suitable to characterize CCs. Therefore, we exploit this relation in order to design a new criterion that quantify the *generality of concepts*.

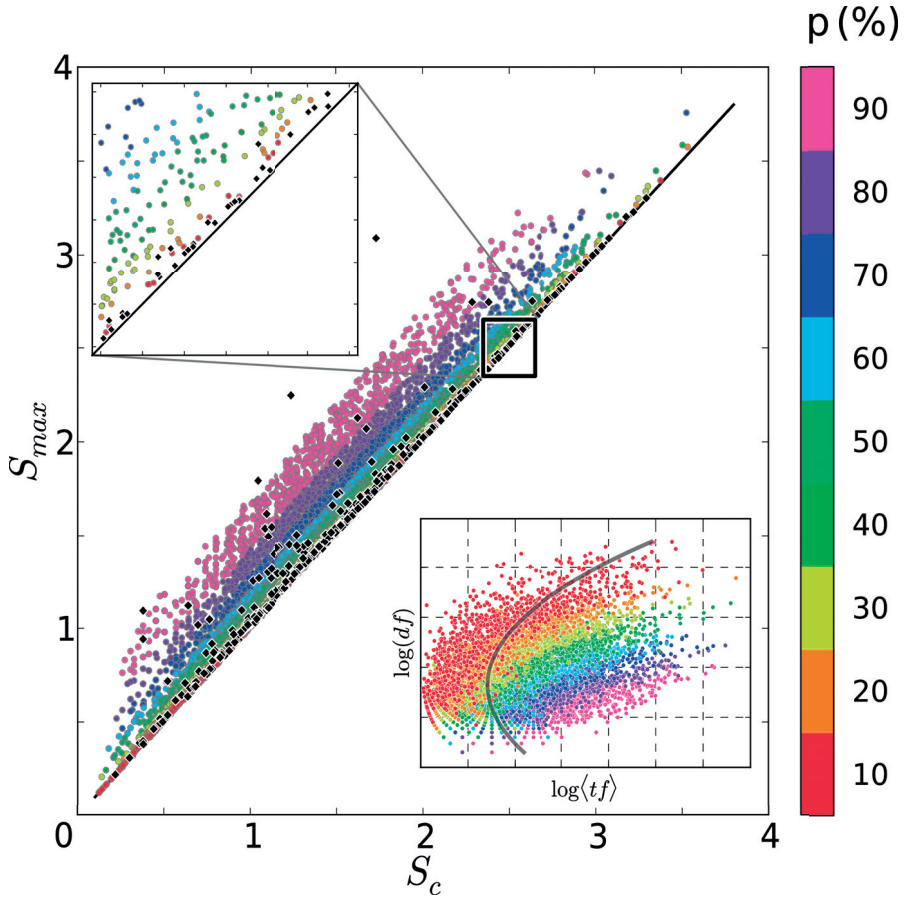

Figure 2.8 – Organization of the concepts in the (S_c, S_{max}) plane. Points are colored according to the percentile p of the residual entropy distribution $P(S_d)$ to which they belong (concepts with $p > 90\%$ are omitted). CCs, represented as black diamonds, have a Pearson correlation coefficient of $r = 0.979$. The solid black line indicates the equality $S_c = S_{max}$. The bottom right inset shows the position of the concepts in the $(\log\langle tf\rangle, \log(df))$ plane, whose color represents the percentile p.

To this aim, we consider for every concept c the *residual entropy*, $S_d(c)$, which is the difference between its maximum and actual entropy, $S_d(c) = S_{max}(c) - S_c(c)$. Albeit this definition is very naive, it is exactly the Kullback-Leibler divergence (also known as relative entropy [185]) between the observed term-frequency distribution, p_c, and the maximum entropy counterpart, q_c,

as shown in subsection A.1.3. A deviation of the actual distribution from the maximally random (with constraints) is therefore an indication that the observed distribution is non-trivial, including peculiar features that are not expected. The tendency of CCs to have a nearly maximal entropy is even more evident from the distribution of S_d displayed in Figure 2.9. The residual entropy distribution of CCs, indicated by the black points, is significantly dissimilar from the one of the other concepts shown by orange dots. Thus, CCs are not just drawn at random from the existing concepts but they feature a very small residual entropy. A small value of S_d for a given concept allows to consider it as generic since it shares this salient property with CCs, meaning that the term-frequency distribution is essentially unbiased apart from respecting the constraints.

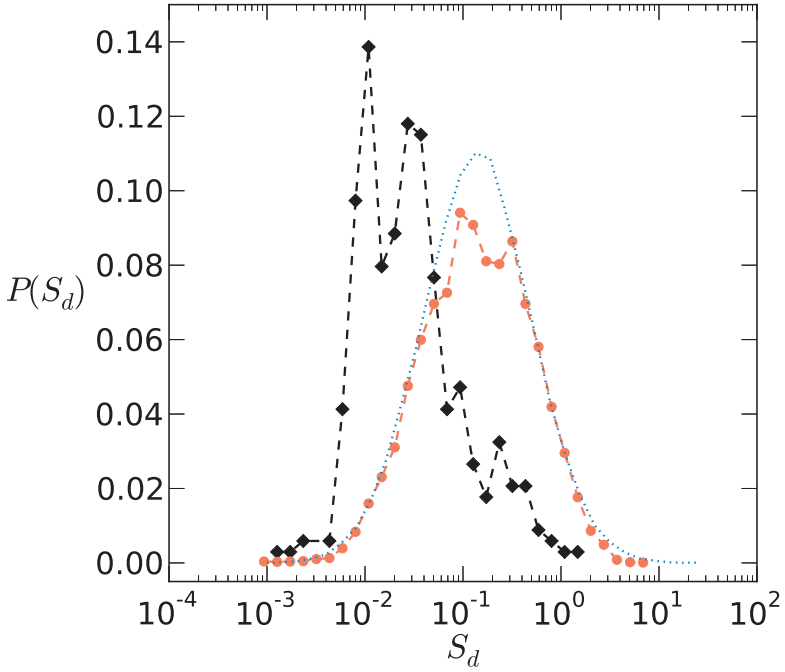

Figure 2.9 – Distribution of the residual entropy $S_d = S_{max} - S_c$ for the CCs (in black) and the other concepts (in orange). The dotted blue line represents the lognormal distribution computed for concepts that were not tagged as common using $\mu = \langle \ln(S_d) \rangle = -1.947$ and $\sigma^2 = \langle \ln(S_d^2) \rangle - \mu^2 = 1.550$ as parameters.

An intuitive explanation of the behavior of a significant concept is then the following: such kind of concept is not so mainstream but rather specific. Therefore, it has not been adopted regularly as other, more generic concepts. As a consequence, the observed term-frequency distribution is somehow deviating from the standard one since its usage is more erratic, responding to a precise need of conveying an accurate meaning. Loosely speaking, significant concepts are not permeating the language, therefore they are used in a way that does not match expected characteristics at the microscopic level of the term-frequency distribution. The residual entropy S_d can be considered as a "distance" from the maximum entropy line $S_c = S_{max}$. Because the distribution of S_d spans different orders of magnitude, imposing a raw value as a threshold to discriminate generic concepts is not appropriate. In order to adapt the classification of generic concepts to the shape of distribution of S_d, we take advantage of the notion of percentile defined

as the value below which a fraction p of the observations fall. The distribution of S_d is then divided in nine percentiles ranging from $p = 10$ to $p = 90$ and concepts are assigned to the percentile p they belong to, as illustrated in Figure 2.8. The color of the points represents the percentile p and the bottom right inset encodes the percentile information in the $(df, \langle tf \rangle)$ plane. Finally, the criterion to establish if a concept is generic is based on the percentiles: concepts that are included in a percentile p are considered *generic* while the rest are *significant*.

A more detailed examination of Figure 2.8 discloses other two intriguing facts. On the one hand, some concepts tagged as common are positioned far away from the diagonal $S_c = S_{max}$, revealing the presence of outliers. Examples of such concepts are 'operational calculus', 'Fraunhofer line', 'gigawatt' and 'Gaussian symplectic ensemble', which may be regarded as generic within some specific field but are fairly specialized for the heterogeneous corpus of article under scrutiny (see Figure 2.1). On the other hand, multiple concepts are located near the diagonal but have not been tagged as common, *e.g.* 'statistics', 'intensity', 'Hamiltonian', 'fluid dynamics', and 'scaling law'. Therefore, we claim that those concepts have been overlooked by the experts without being classified as common despite they are likely so. Both instances demonstrate the drawbacks of the human-based concept tagging, pointing out the benefits of our entropy-based model which is nearly unsupervised.

The novelty of the proposed approach to outline significant concepts stems from the comparison of the observed term-frequency distribution with the theoretical expectation that arises from the maximum entropy principle. Indeed, the statistical characterization of the word distribution has been already addressed in the literature where several models have been proposed to describe the different facets of word usage [224–227, 229, 230, 232]. Given the regularities uncovered in the large scale patterns of word consumption, the idea that they can be explained by a maximum entropy mechanism have been widely employed to construct such models. In parallel, various flavors of the observed entropy of words have been conceived in natural language processing to gauge the importance of words within a text [228, 233–235]. In particular, these studies focus on the analysis of documents composed by different parts (*e.g.* chapters) in order to quantify the entropy of the words based on their arrangement among the parts. The proposed approaches provide the entropy of a word only at the level of individual texts without addressing its relevance for the entire corpus of documents. Moreover, the difference between the maximum entropy and the actual one is not taken into account. On the contrary, such difference is the cornerstone of our method which aims to characterize relevant words that distinguish the composition of documents at the corpus scale. Indeed, the method is also effective to assess the relative performance of concepts in describing the content of a single article, as shown in Tables A.2 – A.4. The adopted approach is powerful enough even if the content of documents is modeled in the simplest way, *i.e.* using the so-called *bag-of-words* approximation where the relative position of words inside the text is not considered.

In such approximation, the raw information about the words frequency across documents may be adopted to define an entropy which is slightly different from the conditional one. Albeit the choice of S_c seems the most straightforward and intuitive, we can construct an entropy based on

the full probability of the term-frequency, p_f, which takes into account also the frequency of no appearance of a concept. This probability distribution is based on the total number of articles in the corpus, N_a, in contrast to $p_c(k) = \frac{N_c(k)}{N_c}$ that includes only the articles $N_c \leq N_a$ where the concept is found. The new term entering in p_f corresponds to the absence of the concept ($k = 0$) and is calculated as the fraction of papers where the concept is not present, $p_f(0) = 1 - df$. The complete probability that the concept appears k times, with $k \in [0, \infty]$, is then $p_f(k) = \frac{N_c(k)}{N_a}$. Therefore, the *full entropy* S_f associated to $p_f(t)$ reads

$$S_f = -\left(1 - df\right) \ln\left(1 - df\right) - df \ln\left(df\right) + df\, S_c\,. \tag{2.17}$$

The interested reader can refer to subsection A.1.1 for the derivation of such expression. Here, we only note that the full entropy features the presence of the df which modulates the contribution of the conditional entropy S_c. In the same fashion we analyzed the relationship between characteristic quantities of the concepts and S_c in Figure 2.5, we inspect the behavior of S_f in order to understand if it exhibits any peculiar trend for CCs with respect to some feature. The comparison of the results shown in Figure 2.10 for S_f with the respective panels of Figure 2.5 demonstrates that neither S_f nor S_c are able to discriminate CCs based on df (black points in panel A). However, the tendency of CCs to have higher entropies for a given value of $\langle tf \rangle$ is observed only in the case of S_c (panel B), the same conclusion being valid for $\langle \ln(tf_c) \rangle$ (panel C). The reader may argue that the above features are very naive and there could be others, more complicated features that are better suited to improve the performance of S_f. Further attempts to detect potential trends of CCs using other features are examined in Figure A.1.

In conclusion, none of the presented combination is useful to isolate CCs. The full entropy S_f is then unfit to highlight CCs and the conditional entropy S_c is, by far, the most suitable and informative quantity. Clearly, we expect that any other naive quantity not based on the notion of information is not able to highlight CCs, as we already discovered for the df in Figure 2.3. Nevertheless, we examine the effects of selecting concepts according to the percentiles of the *inverse document frequency*, IDF, in Figure A.4 since it is a basic feature commonly used to filter out pervasive words in natural language processing. The selection of concepts based on this measure differs from the one of our approach, albeit the two rankings are not completely unrelated.

2.3 Effects of the entropic selection of relevant concepts

The criterion based on the distribution of S_d allows to divide concepts for different percentiles. Increasing the percentile p, only meaningful concepts (below p) that incorporate significant information are retained to construct network of similarity between articles while generic concepts (above p) are discarded. In this way, we finally achieve the desired *ex-ante* approach to filter concepts. The outcomes of the pruning on the topology of the networks are reported in Table 2.2. The total number of concepts, $N_{con} = |\mathcal{C}|$, as well as the number of documents containing at least one concept, N_a, diminish as p increases, although the latter remains fairly constant up to

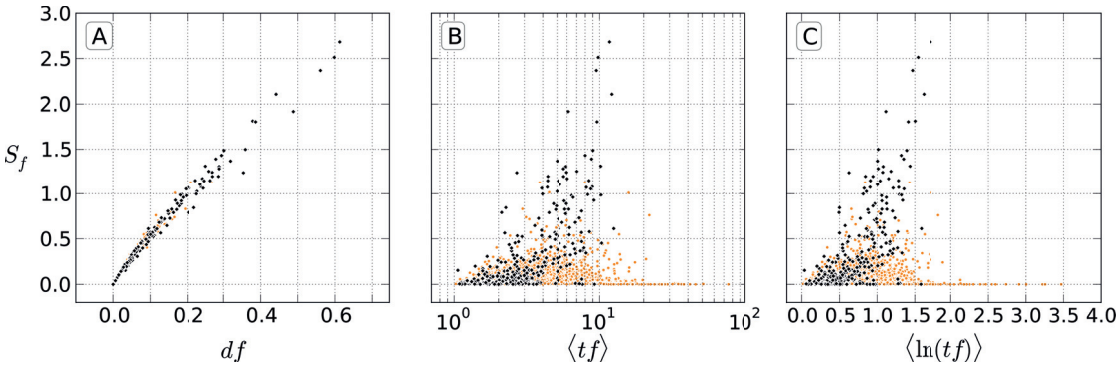

Figure 2.10 – Relations between the full entropy S_f and several features of the concepts. Common concepts in SW (CCs) are denoted by black points. The different panels display the relation between S_f and the fraction of articles where concepts appear in, df (A), the average term-frequency, $\langle tf \rangle$ (B), and the average logarithm of the term-frequency, $\langle \ln(tf) \rangle$ (C).

$p = 30\%$. The link density ρ, instead, undergoes a sizable drop from 36% to 7% when p jumps from 0% (no filtering) to 10%. As a consequence, both the maximum and average degrees, k_{max} and $\langle k \rangle$, dramatically decrease while the average "distance" between articles, $\langle l \rangle$, grows with p. At the global level, the connectedness of the networks gets attenuated as outlined by the increment of $\langle l \rangle$ with p and the concurrent fragmentation into separate connected components M. This effect is the byproduct of the presence of "*cultural holes*" among different domains in physics already highlighted for scientific disciplines [211]. Cultural holes, indeed, quantify the difficulty in communicating between disciplines from an information-theoretic measure of the diversity in the respective languages.

As the pattern of connections becomes sparser, spurious similarities tend to fade. Ergo, articles on the same theme are more similar and develop stronger relations with each other. Groups of tightly connected articles then emerge more neatly from the network structure leading to a refined organization in communities. In order to discover which communities are present in a network, we maximize the weighted modularity by means of the Louvain method described in subsection 1.1.4.

2.3.1 Organization of articles into topics

Articles sharing similar concepts – *i.e.* on the same topic – tend to belong to the same community. In terms of the community structure, we can imagine that the fraction of pruned concepts, p, acts as a parameter, tuning somewhat the granularity of the topics. Therefore, we investigate how the community structure evolves as a consequence of the entropic filtering. For each percentile p, we executed the Louvain algorithm 1000 times using a different random seed node per run. The variability in the number of detected communities for each run is then displayed in Figure 2.11. Augmenting the filter intensity p there is a remarkable tendency toward an increased number of communities. This effect is clearly due to pruning of the networks in terms of the link density

p (%)	N_{con}	N_a	ρ (%)	$\langle k \rangle$	k_{max}	T	$\langle l \rangle$	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	4264	34939	0.482	168.307	1999	0.508	2.914	20
70	3197	24710	0.363	89.766	1140	0.755	3.409	59
80	2132	14789	0.257	37.989	495	0.783	4.242	153
90	1066	5703	0.228	13.027	104	0.848	7.124	342

Table 2.2 – Topological quantities of the article similarity networks. The first row ($p = 0\%$) corresponds to the original (unfiltered) network, while the following to the networks obtained after removing generic concepts at a given percentile p. The columns indicate the percentage of filtered concepts p, the number of concepts N_{con}, the number of articles (nodes) containing at least one concept N_a, the link density ρ (Equation 1.3), the average and maximum degrees, $\langle k \rangle$ and k_{max} (Equation 1.4 and Equation 1.5), the transitivity T (Equation 1.7), the average shortest path length $\langle l \rangle$ (Equation 1.10) and the number of connected components M (see subsection 1.1.3). The minimum edge weight is equal to $w_{min} = 0.01$.

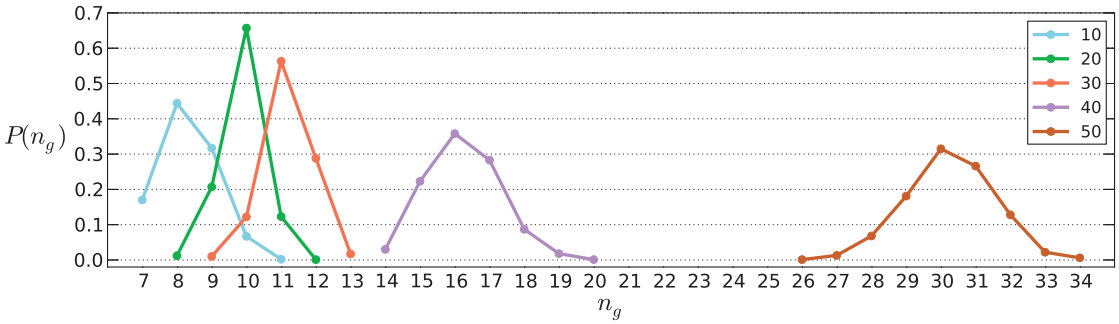

Figure 2.11 – Probability of the number of discovered communities, $P(n_g)$, over 1000 runs of the Louvain algorithm for different filtering intensities p denoted by different colors in the legend.

and the breaking of the networks in distinct connect components outlined in Table 2.2. Other statistics involving the similarity between the partitions detected in different runs are reported in Figure A.2.

Focusing on the partition with the highest modularity we investigate its composition in terms of concepts in order to understand the effects of the filtering at the community level. The results are displayed through the Sankey diagram[7] [237] in Figure 2.12. Each community is identified by a box whose height represents the community size, while each column refers to a filter intensity p. The name of a community describes its main topic identified from the ten most used concepts among its articles. For $p = 0$, the communities are clearly associated to major domains in physics.

[7]The interactive version of the diagrams displaying additional information is available at [236]

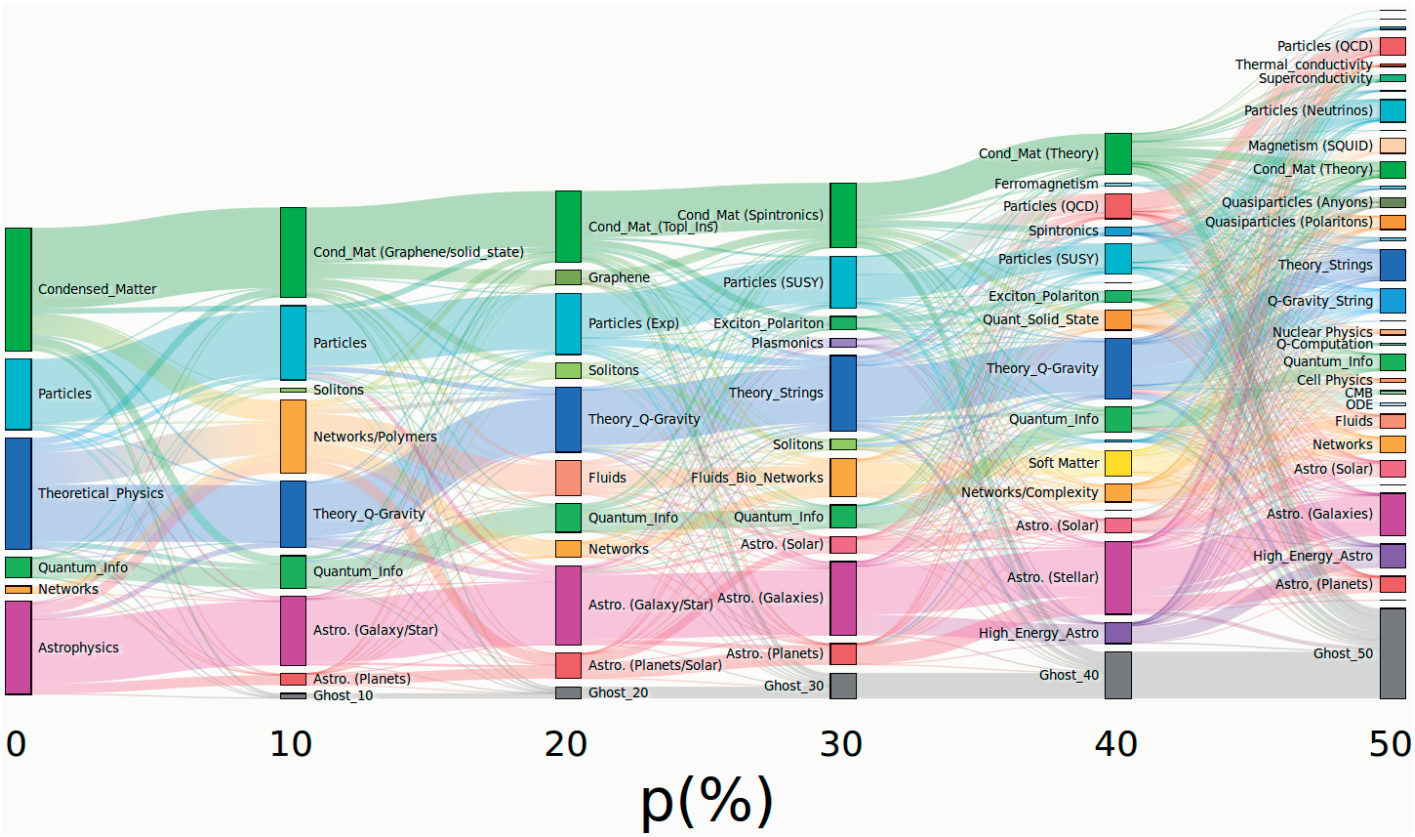

Figure 2.12 – Static Sankey diagram of the physics dataset. Each community is illustrated as a colored box whose height is proportional to the number of papers it contains. The label close to the box represents the topic assigned to the community considering the ten most frequent concepts within, *i.e.* those that are present in more articles. Different shades of a color are used to highlight similar or related topics. The gray box, labeled as "ghost", contain papers that do not have any connection with others since they are only composed by generic concepts that have been deleted. The width of the bands running between boxes corresponds to the number of shared articles between communities. Each column denotes the community structure at a different filter intensity p which identifies the percentage of generic concepts that have been removed. The partition shown for every value of p is the one with the best weighted modularity over 1000 runs of the Louvain algorithm starting from different initial nodes, as explained in subsection 1.1.4. The interactive version showing additional information is available at [236].

At this level, it is not possible[8] to detect finer-grained communities because of the effect of generic concepts that hold together the articles within large groups. As p increases, a progressive fragmentation of the topics takes place, moving from broad domains in physics – not exactly overlapping with the `arXiv` classification as shown by [238] – to more specialized themes at $p = 20\%$. A paradigmatic example of the fragmentation is observed for the Astrophysics community at $p = 0\%$ which constitutes a major field that progressively unfold into Stellar Physics, Planetary Astrophysics, High Energy Astrophysics and Solar Physics up to $p = 40\%$.

Although the filtering of generic concepts allows to unveil smaller and more precise communities, it causes the loss of information that is encoded in the concepts. Therefore, we do not want to discard too many concepts. Indeed, considering large values of p deteriorates the results. Actually, this is the consequence of two combined effects. On the one hand, the eradication of too many generic concepts (which may be, in part, mildly generic) makes the similarities between articles depending on concepts that are extremely specific, whose statistical relevance is weak and that are vaguely related to each other. On the other hand, an increasing portion of articles containing only generic concepts are no longer part of the network as they do not share any similarity with other articles. These isolated articles (*i.e.* without significant concepts) constitute the so-called "ghost" community. An heuristic rule to establish the level p_{opt} at which to stop the filter is given by the size of the "ghost" community that should not exceed the average size of the other communities, as illustrated in Figure 2.13. In this case, $30\% \leq p_{opt} \leq 40\%$.

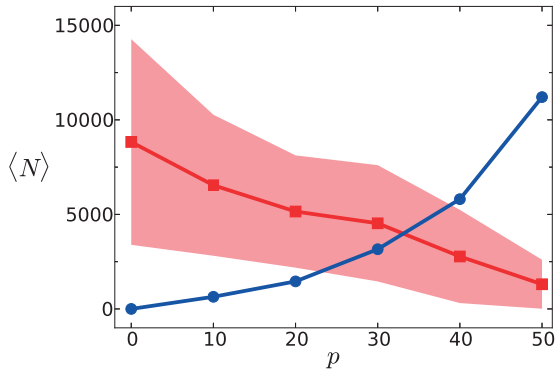

Figure 2.13 – Average community size, $\langle N \rangle$, as a function of the filter intensity p for the networks of articles. Red circles refer to the size of the "ghost" community, while blue squares denote the average size of all the other communities. The shaded area corresponds to the standard deviation of the community size.

[8] To be precise, the assertion is true if we fix the community detection method (modularity maximization) and the algorithm (Louvain). However, as we mentioned in subsection 1.1.4, many other methods are available that eventually discover hierarchies of communities. If we may want to adopt any of them to highlight smaller communities, their potential is limited nevertheless since the similarities encode some noise due to generic concepts that create spurious connections and increase the weight of links between nodes. Therefore, any method would be applied to a pattern of similarities which is not "clean", undermining the advantages of such methods.

2.3.2 Filtering keywords in web documents

In principle, the residual entropy approach described above should be effective not only in discriminating concepts in scientific articles but also in detecting generic keywords in any kind of document. To assess the validity of the method in another context, we repeat the same analysis detailed in this Chapter for the corpus of physics articles from `arXiv` on another corpus of web documents about climate change. Before delving into this dataset it is worth mentioning that, in general, a web document lacks the same structural organization of a scientific manuscript. Indeed, there is not an introduction, a conclusion, or a methodological part, for example. This fact is linked to the different perspective that an online document aims to convey with respect to a scientific paper even if on the same topic, as they can simply report facts (news releases) or provide an opinion on a given subject.

Our corpus of web documents has been constructed from a collection of tweets available within the ScienceWISE database (SW). More specifically, such collection contain tweets on climate change posted between January and June 2015 and it was compiled using the Twitter API [239] through several harvesting campaigns. Among these tweets, we considered only the *original* ones, thus culling mentions, re-tweets and other similar "non original" posts. Then, we kept only tweets written in English containing at least one URL. Such procedure gave a set of distinct URLs pointing to some fifty million web documents ranked according to the number of tweets mentioning each of them. At the end, we took the 100.000 most "tweeted" URLs and we retained only those that contain at least one of 165 *specific* keywords on climate change in order to ensure a thematic consistency.

From this procedure we obtained a corpus of 30705 documents. To get rid of very short documents, we removed from the corpus those with less than $L_{\min} = 500$ words that roughly correspond to half of a book page. After this thresholding, we ended up with a corpus of 18770 documents. Since the SW platform does not have a curated ontology of crowdsourced concepts on climate change, we resorted to mine simple **keywords** composed by n-grams from the texts using the KPEX tool as it is natively implemented in the SW platform [219]. The KPEX algorithm returned 822.601 unique keywords which were stemmed first and then lemmatized, obtaining a final set of 152.871 keywords. Clearly, keywords are more "rough" in some sense as we cannot expect that they have a precise meaning and a high specificity like concepts. Nevertheless, we will refer to them as concepts in the following. Another important distinction between web documents and scientific articles concerns their length. The comparison of distribution of the number of words per document between the two corpus reveals that they have different traits, as displayed in Figure 2.14. Specifically, in the climate change corpus the distribution of the number of words per document does not to possess a characteristic scale and is quite inhomogeneous. Because of this, taking into account only the raw term-frequency when defining the entropy is not ideal. Nevertheless, the same approach to classify keywords can be applied with a minor change, replacing the term-frequency of a concept c in document α, $tf_c(\alpha)$, with its *term-frequency density*, $rtf_c(\alpha) = \frac{tf_c(\alpha)}{L(\alpha)}$, where $L(\alpha)$ is the length of the document in terms of the number of words. As a consequence, rtf_c is a continuous variable and the entropies must be redefined by

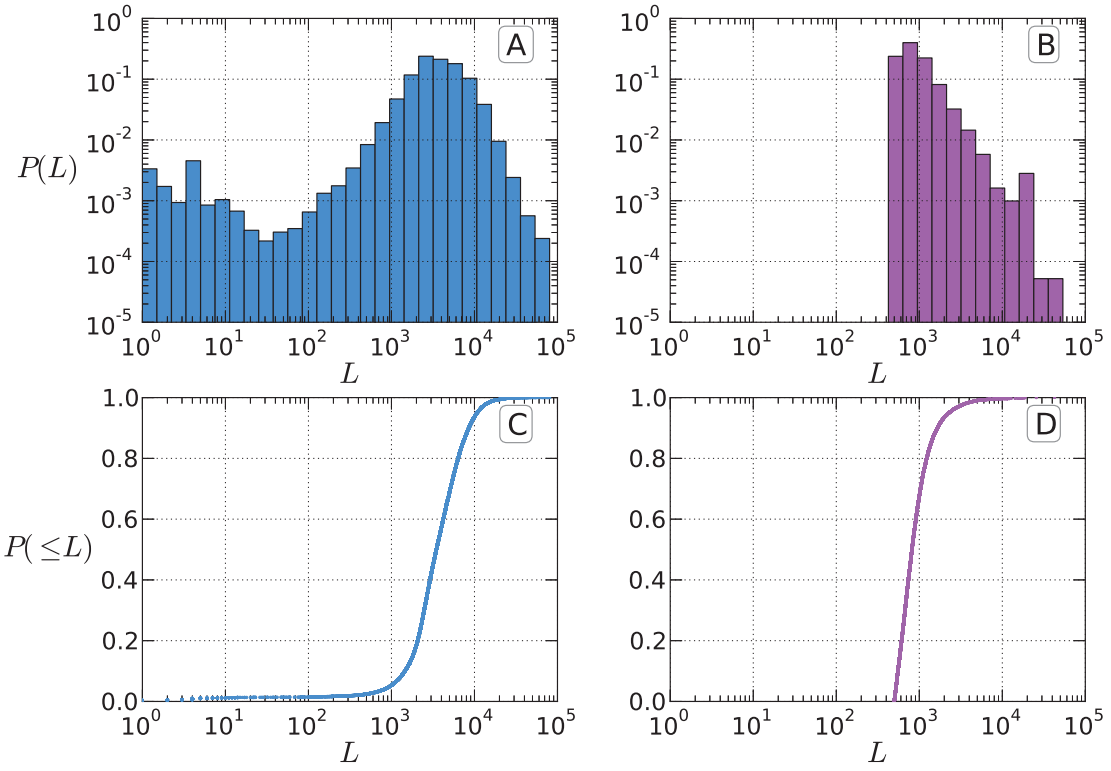

Figure 2.14 – Distribution $P(L)$ of the number of words per document, L, for the physics (A) and climate change (B) corpus. In the first case a clear peak is present, indicating a typical length of the documents around $L \simeq 3000$, while this is not the case for climate change. Panels (C) and (D) display the cumulative distribution functions $P(\leq L)$ for the same corpora.

substituting the summation with an integral over the probability density function $p_c(x)$ of rtf_c. Moreover, keywords for which $\max(rtf_c) - \min(rtf_c) < 0.005$ are ignored in order to limit those with very similar values of rtf_c. As a result, the number of keywords gets shrunk to 9222.

The maximum entropy probability density function is recovered based on two constraints, average and variance of the logarithm of the term-frequency density, $\langle \ln(rtf_c) \rangle$ and $\sigma^2(\ln(rtf_c))$. We select 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 $p_c(x)$ satisfying these constraints is a lognormal, *i.e.* the distribution associated to a variable x whose logarithm, $y = \ln(x)$, is normally distributed:

$$p_c(x) = \frac{1}{\sqrt{2\pi}\,\sigma\,x} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] \qquad \text{with } x > 0\,. \tag{2.18}$$

The parameters μ and σ^2 that appear in can be directly calculated from the observed data:

$$\mu = \int_0^\infty \ln(x)\, p_c(x)\, dx \equiv \langle \ln(rtf_c) \rangle\,, \quad \sigma^2 = \int_0^\infty \big(\ln(x) - \mu\big)^2\, p_c(x)\, dx \equiv \sigma^2(\ln(rtf_c))\,. \tag{2.19}$$

The maximum entropy S_{max} associated to the probability in Equation 2.18 is then:

$$S_{max} = -\int_0^\infty p_c(x) \ln p_c(x)\, dx = \ln\left(\sqrt{2\pi}\,\sigma\right) + \mu + \frac{1}{2}. \tag{2.20}$$

The reader can find further details about the methodology in subsubsection A.1.2.2, while in Figure A.6 we show the relation between the conditional entropy, S_c, and the maximum one, S_{max}, for the keywords. Filtering keywords based on the percentile p of the residual entropy, $S_d = S_{max} - S_c$, we recover sparser similarity networks whose topological quantities are reported in Table A.5. The community structure of the networks in response to the selective removal of concepts highlights the progressive specialization of topics, as displayed in the Sankey diagram of Figure 2.16. The comparison with the results for the physics corpus in Figure 2.12 indicates one significant difference: the communities about *extreme weather/energy storage* tends to progressively condensate going from $p = 5\%$ to $p = 20\%$. To gain insight into such phenomenon, we focus on the set, $\widetilde{\mathcal{C}}$, of 20 most frequent keywords in each community s. Then, we compute its *coverage*, $\Gamma_s(\widetilde{\mathcal{C}})$, defined as the union of the sets of documents where those keywords appear divided by the size of the community, *i.e.* the number of documents N_a^s. Hence, $\Gamma_s(\widetilde{\mathcal{C}}) = \frac{1}{N_a^s} \cup_{c\in\widetilde{\mathcal{C}}} N_a^s(c) \in [0,1]$. Remarkably, the coverage of the community named "Mixed_themes" at $p = 20$ 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 keywords in community "Mixed_themes" indicates that documents condense into a single community due to the similarities associated to small groups of keywords weakly related together. However, the deep reason behind such condensation is the presence of keywords whose distribution is not well-describe by a lognormal, as shown in Figure 2.15 for the 10 most frequent ones. The discrepancy between the observed distributions and the lognormal fits is the motivation at the basis of the limited ability to characterize the content of the web documents in the community. When we take off these keywords, the similarities joining together this vast condensed community dissolve – or become weaker, at least – thereby fragmenting it into much smaller communities that address more specific topics.

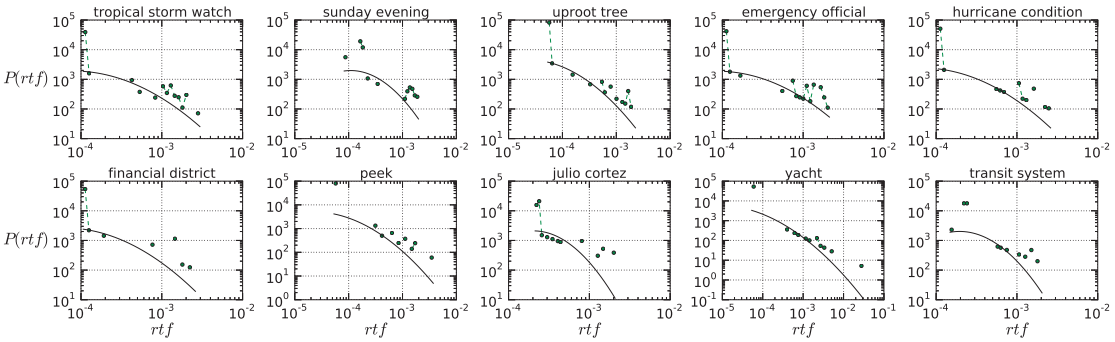

Figure 2.15 – Distribution of the top ten most frequent concepts within the community of uncertain label "Mixed_themes" found at $p = 20\%$ in Figure 2.16. The lognormal fit of each distribution is plotted with a black line. All the distributions deviate considerably from their lognormal fit, meaning that the Kullback-Leibler distance from the lognormal distribution to the observed one is high.

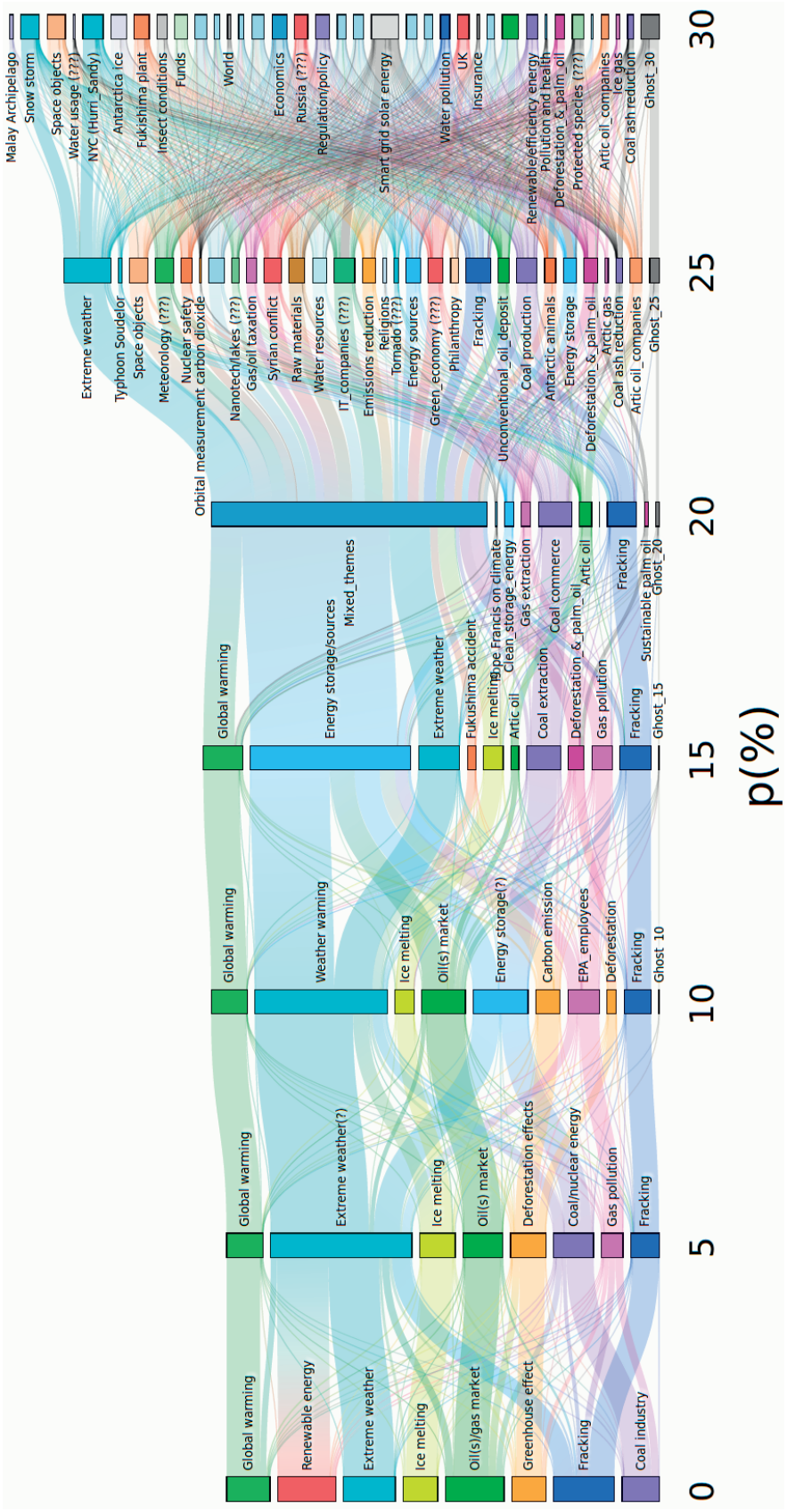

Figure 2.16 – Static Sankey diagram of the climate change collection. Each community is represented as a colored box whose height is proportional to the number of web documents it contains. A topic is assigned to each box according to the 10 most used keywords, *i.e.* those appearing in more papers. The thickness of the bands between boxes indicates the number of shared web documents. Each column denotes a different intensity of filtering p. Concepts are pruned according to their residual entropy S_d computed from rtf. The minimum fluctuation of rtf is equal to 0.005. Interactive version available at [236].

Finally, we establish the optimal filtering level p_{opt} with the same heuristic previously adopted. Therefore, p_{opt} is estimated as the crossover between the size of the "ghost" community and the average size of the other communities, as illustrated in Figure 2.17. In this case, $25\% \leq p_{opt} \leq 30\%$.

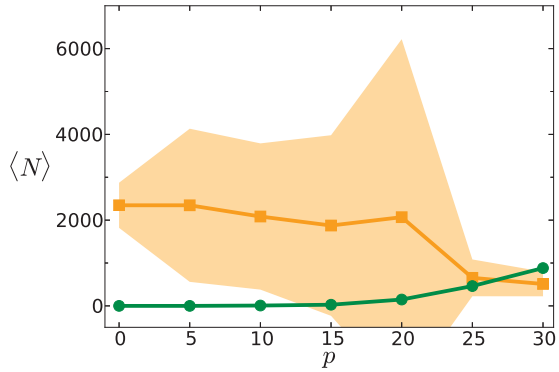

Figure 2.17 – Average community size, $\langle N \rangle$, as a function of the filter intensity p for the web document networks. Green circles refer to the size of the "ghost" community, while yellow squares denote the average size of all the other communities. The shaded area corresponds to the standard deviation of the community size.

Generally speaking, the investigation of the unfolding of the community structure varying the filtering intensity shows three different behaviors:

I) preservation with specialization, *i.e.* when the topic of a community remains unaltered but the concepts used to characterize it are more specific (*e.g.* Cond_Mat/Astrophysics in Figure 2.12);

II) splitting with specialization, *i.e.* when the elimination of generic concepts causes the fragmentation of the original topic into more specific sub-topics (*e.g.* Cond_Mat → Graphene + Solid State in Figure 2.12);

III) nucleation (*e.g.* Extreme weather in Figure 2.16) *i.e.* when one or more shared concepts draw documents together into a single community.

Concluding, it is worth to underline two major differences between the climate change corpus and the physics one. First, there is no guarantee that web documents comes from trusted sources (like newspapers online editions or news websites) while articles in `arXiv` cannot be uploaded by someone that is not endorsed by an author that already published on `arXiv` . Further, the text of articles is parsed to spot those using a very atypical jargon which likely refers to research themes out of the conventional scientific landscape, perhaps containing more delirious visions than rational claims[9]. Therefore, the classification scheme of `arXiv` article can be used to

[9] See footnote number 2 of p. 223 in [211].

roughly understand the composition of the corpus in terms of domains. On the contrary, such thematic division is not present for web documents. Second, keywords extracted from web documents are less precise than scientific concepts, implying that the similarities between web documents are less accurate. This notwithstanding, the entropic method to filter generic keywords is effective in order to construct similarity networks between documents that are sparser, encoding weaker spurious interactions between documents. As a consequence, these networks are more computationally tractable for community detection techniques.

Nevertheless, examining the community structure of the article similarity networks is not the unique way to understand the effects of the concept filtering on the dataset under scrutiny. In order to gain insights into the composition of a corpus, a complementary approach is to model it as constituted by documents containing a mixture of topics, each characterized by specific words. Such kind of analysis, called *topic modeling*, is indeed the mainstream approach that is considered with the aim of exploring the themes inside a corpus [71, 72]. In the following part, we are going to dissect the results of the topic modeling on the corpus of physics article, comparing them with the communities of articles previously described.

2.3.3 Topic modeling

The goal of *topic modeling* is the description of the thematic structure of a document corpus. To accomplish this objective, several techniques have been developed to automatically learn the composition of topics in terms of words describing a corpus [71, 72, 240]. The first attempts were based on the idea that subjects can be identified from a decomposition of the word-document matrix $\boldsymbol{M}$ though dimensionality reduction techniques [208]. Given the $W \times D$ matrix $\boldsymbol{M}$, whose rows represent words, and its columns represent documents; Latent Semantic Analysis (LSA) [241] – also known in information retrieval as Latent Semantic Indexing (LSI) [242, 243] – is based on calculating the Singular Value Decomposition (SVD) [244] of $\boldsymbol{M}$ as follows:

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T = \sum_{r=1}^{R} \lambda_r \boldsymbol{u}_r \cdot \boldsymbol{v}_r^T, \tag{2.21}$$

where R is the rank of $\boldsymbol{M}$ and $\boldsymbol{\Lambda}$ is the diagonal, $R \times R$ matrix whose elements λ_r are the singular values of $\boldsymbol{M}$, *i.e.* the eigenvalues of $\boldsymbol{M}\boldsymbol{M}^T$ such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_R$. The matrix $\boldsymbol{U}$ is a $W \times R$ matrix with orthonormal columns $\boldsymbol{u}_r$ and $\boldsymbol{V}$ is a $D \times R$ matrix with orthonormal columns $\boldsymbol{v}_r$. In LSA, only the K largest singular values of the decomposition are retained (usually $K \ll R$). The approximated version of $\boldsymbol{M}$, $\boldsymbol{M}_K = \boldsymbol{U}_K\boldsymbol{\Lambda}_K\boldsymbol{V}_K^T$, is a matrix of rank K where the rows $\boldsymbol{V}_K\boldsymbol{\Lambda}_K$ are a compressed representation of documents in the K-dimensional space described by the columns of $\boldsymbol{U}_K$. For a given K, the matrix $\boldsymbol{M}_K$ is then taken among the matrices G of rank at most K as those that minimizes the reconstruction error quantified by the Frobenius norm $\left\|\boldsymbol{M} - \boldsymbol{G}\right\|^2 = \sum_{w,d} (\boldsymbol{M}_{wd} - \boldsymbol{G}_{wd})^2$. Another dimensionality reduction technique, called Non-negative Matrix Factorization (NMF) [245], have also been applied with the same purpose of providing a compressed representation of the word-document matrix, $\boldsymbol{M}$, which eventually captures the thematic features of a corpus through the out-of-the-box clustering of the columns of

$\boldsymbol{M}$ [246, 247]. In such case, the word-document matrix $\boldsymbol{M}$ is approximated as $\boldsymbol{M} \approx \boldsymbol{Z}\boldsymbol{H}$ where $\boldsymbol{H}$ is a $K \times D$ non-negative matrix whose non-zero entries in the kth row indicate the membership of the documents in $\boldsymbol{M}$. On the other hand, $\boldsymbol{Z}$ is a $W \times K$ non-negative matrix whose entries in the kth column represent the centroid of the kth component that describes a thematic feature. However, using LSA has some disadvantages [248]: for example, the reduced dimensionality K transforms the original sparse matrix $\boldsymbol{M}$, into an approximate $\boldsymbol{M}_K$ whose factors $\boldsymbol{U}$ and $\boldsymbol{V}^T$ are dense. Therefore, it is not straightforward to associate words or documents exclusively to a meaningful subsets of the K components. Moreover, the model cannot capture non-linear dependencies since it is based on linear algebra. To overcome such limitations, a full probabilistic model has been introduced as an improvement of LSA to describe the presence of words into documents by means of the association of words to topics and topics to documents. Such model is called probabilistic Latent Semantic Analysis (pLSA) [249, 250], a Bayesian approach for the unsupervised discovery of the hidden topics in a corpus. Interestingly, there are some analogies between NMF and pLSA methods [251]. In particular, when the error function to minimize in order to find the approximated representation of $\boldsymbol{M}$ is the Kullback-Leibler divergence [185, 186] (see Equation 1.36), NMF is equivalent to pLSA. Nevertheless, strictly speaking, a topic model is any *probabilistic* generative model that describes a corpus in terms of topics.

The basic ingredients of such a model are the probability that a given document d contains a topic t, $\pi(t|d)$, the probability that a word w belongs to a topic t, $\pi(w|t)$, and the probability of the given document, $\pi(d)$. These probabilities are estimated by maximizing the likelihood $\mathcal{L}$ of generating the observed distribution of words in documents, $\pi(w, d)$:

$$\mathcal{L} = \prod_{w,d} \pi(w, d) = \prod_{w,d} \sum_{t} \pi(w|t)\pi(t|d)\pi(d). \tag{2.22}$$

In general, for a corpus of D documents with W unique words and T topics, we need to estimate $D \times T$ probabilities $\pi(t|d)$ and $T \times W$ probabilities $\pi(w|t)$. In pLSA all these probabilities are directly treated as parameters, therefore they must be estimated from the observations. This plethora of parameters, however, is not ideal to extract latent topics since there are all independent, thereby assuming the absence of any constraint between topics which is unrealistic. In particular, the number of parameters increases linearly with the number of documents D, hence the model is not fully generative. Instead, a better approach would be to add some regularity among such parameters. To this purpose a generalization of pLSA, called latent Dirichlet allocation (LDA) [68, 252], has been developed. LDA differs from PLSA since the probability distribution over the T topics of a document d, $\boldsymbol{\Pi}(T|d) = \{\pi(t|d)\}_{t=1}^{T}$, as well as the probability distribution over the W words in a topic t, $\boldsymbol{\Pi}(W|t) = \{\pi(w|t)\}_{w=1}^{W}$, are drawn from a Dirichlet distribution with T and W parameters, respectively; as such, probabilities are constrained and no longer free parameters. Due to this property, LDA is a fully generative model of documents [253].

Several extensions relaxing the constraints of LDA [254–256], for example describing hierarchical topic modeling [257] or incorporating other information about articles have been introduced (see references in [258, 259]). A further refinement of LDA, called Correlated Topic Model

(CTM) [260] allows to account for the correlation between the occurrence of topics. Indeed, the mixture of topics within a document is modeled by a Dirichlet distribution that implicitly generates independent probabilities for the mixture. For CTM, the topic mixture is expressed by a logistic normal distribution that encompass the possibility of a correlation between topics, thereby allowing a more faithful description of the thematic structure. Nevertheless, the state-of-the-art procedure to uncover the thematic structure of a corpus remains LDA. For such reason, we have decided to use LDA to extract the topic organization of our copora.

From the operational viewpoint, establishing the total number of topics T as well as the composition of each of them are then the crucial objectives of topic modeling. Each topic is composed by a subset of the total number of words W which are semantically related, while each document is described by a small number of prominent topics that contain most of the words that appear in it. On the one hand few topics may tend to aggregate words which are not equally related but, on the other hand, many topics may overfit the model incorporating only very specific words. A remarkable improvement of the LDA performance that does not modify the algorithm itself has been obtained in [69] where the authors borrowed ideas and methods from network theory to successfully guess the composition and number of topics. Here, we briefly describe the proposed pipeline called TopicMapping (TM), which consists of four steps:

Preprocessing documents' text Given a corpus of D documents, preprocessing the texts is a standard stage in text mining [208] which consists in removing the so-called "stop words", namely terms that are syntactic elements and do not contribute to the definition of topics. The remaining words are then stemmed in order to identify their common root like in the case of the singular and plural forms of the same noun or the different tenses of the same verb.

Filtering the similarity between words The presence of the words within documents can be modeled as a weighted bipartite network which involves words and documents as distinct types of nodes. The link weight, a_{ud}, between word u and document d is *term-frequency* of the word in the document, $tf_u(d)$. Instead of evaluating the unipartite projection onto documents (as done before), we define the unipartite network of words, as described in subsection 1.1.2, where the similarity between any pair of words u and v is the dot product of their term frequencies over the documents where they coappear:

$$w_{uv} = \sum_{d=1}^{D} tf_u(d)\, tf_v(d) = \boldsymbol{a}_{u:} \cdot \boldsymbol{a}_{v:}. \tag{2.23}$$

Following this definition, very frequent words are strongly related to any other, biasing the similarity also with more specific words that are semantically closer to each other. In order to quantify this effect, a simple null model is introduced where the words are randomly shuffled among documents preserving the number of words per document. The purpose of the second step of TM is to prune the links between words which can be explained by the coappearance of words by chance. The expected value $\langle w_{ab} \rangle$ for the null model

depends only on the product between the number of occurrences of the two words u and v in the whole corpus, namely $tf_u = \sum_d tf_u(d)$ and $tf_v = \sum_d tf_v(d)$, respectively. The probability distribution of the dot product similarity under the null model is well described by a Poisson distribution $\pi_{\langle w_{ab}\rangle}(k) = \frac{\langle w_{ab}\rangle^k \exp(-\langle w_{ab}\rangle)}{k!}$ with average $\langle w_{ab}\rangle$. The actual similarity between a pair of words is then compared against the expectation for the null model: if the two values are significantly different, it means that the actual tie is unlikely to be attained from the null model, therefore the link is retained in the network of words since it stems from a genuine similarity.

Defining topics From the structure of the filtered network between words, topics are identified as communities of words using the Infomap algorithm [156]. In this way, the number of topics is automatically identified as a result of the algorithm, without the need to fix it *a priori* as required in topic modeling. Topics defined by Infomap are "hard" communities of words, *i.e.* each word is assigned to exactly one cluster. The same word, however, may be potentially used in multiple topics. In order to relax the single membership of words, topics are refined through the application of a standard topic modeling.

Estimating the topic model The detected communities are adopted as initial guess for the number of topics and the word composition of each topic in PLSA. A local maximization of the likelihood is carried out in such a way that the same word is allowed to appear in several topics and documents are mainly described by fewer topic. The model probabilities are then estimated and taken as initial guess for the evaluation of the LDA parameters.

Among the above steps, the pruning of links contributes to denoise the similarity network between words as it removes ties that can be explained by the null model. From this purified network, the communities of words are considered as educated guess for the topics composition which is further refined with LDA, whose results are better than standalone LDA since the parameter optimization depends on the quality of the initial topics. The introduction of the network perspective in topic modeling is then crucial to overcome the performance limits of standard algorithms like LDA.

Provided that TM is the "gold standard" to identify topics in a collection of documents, we apply it to the dataset of scientific concepts in physics articles, exploring the effects of filtering generic concepts at various percentiles p on topic modeling. Each topic t is characterized by the number of concepts it contains, $n_w(t)$, and the probability associated to it, $\pi(t) = \sum_w \pi(t|w)\,\pi(w)$, which indicates its importance. The relation between the two quantities is shown in Figure 2.18 (a): each circle corresponds to a topic uncovered at a given filtering percentile as represented by colors. The number of topics increases as a direct consequence of the filtering since the similarity network between concepts becomes sparser and more communities are detected. Simultaneously, the probabilities $\pi(t)$ associated to the topics decrease, as well as the number of concepts within the topics, $n_w(t)$. Interestingly, topics tend to be more concentrated on a vertical region with low value of $\pi(t)$ spanning almost two orders of magnitude with respect to $n_w(t)$. In the inset of Figure 2.18 (a), we better appreciate that the total number of topics grows fast but the number of topics with a significant probability (*e.g.* , $\pi(t) > 0.01$) does not increase equally rapidly. Such

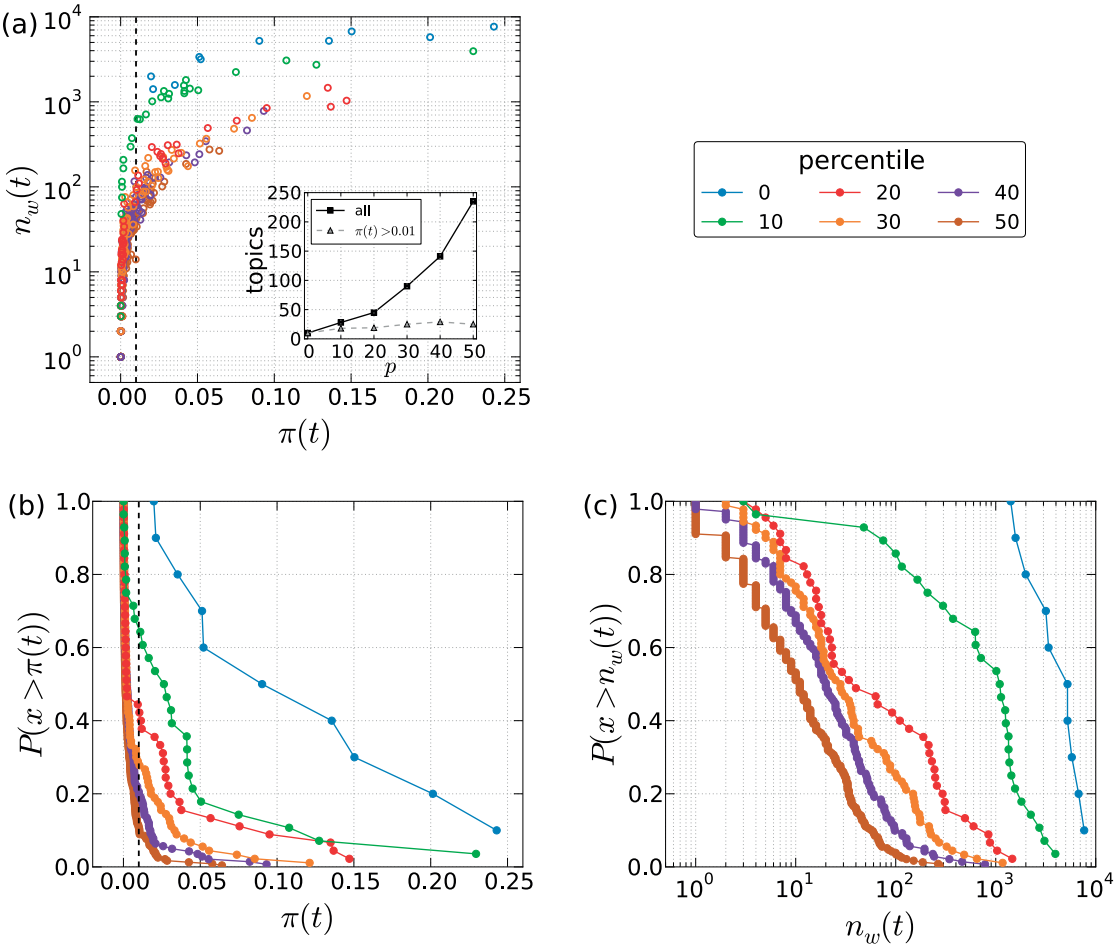

Figure 2.18 – (a) Relation between the number of concepts $n_w(t)$ and the probability $\pi(t)$ associated to each topic t. Every circle represents a topic whose color denotes the filtering percentile. The dashed vertical line corresponds to the topic probability $\pi(t) = 0.01$ below which topics are not considered meaningful. In the inset, the total number of topics for each percentile is shown by squares along with the number of important topics with probability $\pi(t) > 0.01$ shown by triangles. The complementary cumulative distribution functions of the topic probability $\pi(t)$ and the number of concepts per topic $n_w(t)$ are represented in (b) and (c), respectively. Colors identify the percentile of filtered concepts for which the distributions are displayed.

trend of the probabilities $\pi(t)$ varying the percentile of filtered concepts is well illustrated by the complementary cumulative distribution functions (ccdfs) in Figure 2.18 (b) defined as the total fraction of topics with a probability greater than $\pi(t)$. The proliferation of concepts with low probabilities is accentuated especially when passing from $p = 0$ (no filtering) to $p = 20$ as the distributions tend to span a narrower range. Likewise, the ccdfs of the number of concepts within topics, $n_w(t)$, are displayed in Figure 2.18 (c) for various percentiles. Without filtering the concepts ($p = 0$), all the topics are composed by more than 1000 concepts. However, already at $p = 10$ the number of concepts decreases significantly, with roughly half of the topics having less than 1000 concepts. A considerable reduction in the number of concepts takes place at

$p = 20$ where only the 40% of topics have at least 100 concepts. For the successive percentiles the number of concepts per topic are further reduced but the distributions becomes less heterogeneous as more topics have a similar number of concepts.

Once we have provided an overview of the statistics about the topics, we then want to understand how these topics describe the content of the articles by means of the probability of a topic t in a given article d, $\pi(t|d)$. Thus, for every article we extract the maximum of these probabilities over the topics T, $m(d) = \max_{t \in T}(\pi(t|d))$, and compute the ccdf of such values for all the articles as illustrated in Figure 2.19 (a). The distributions reveal that the majority of articles have a

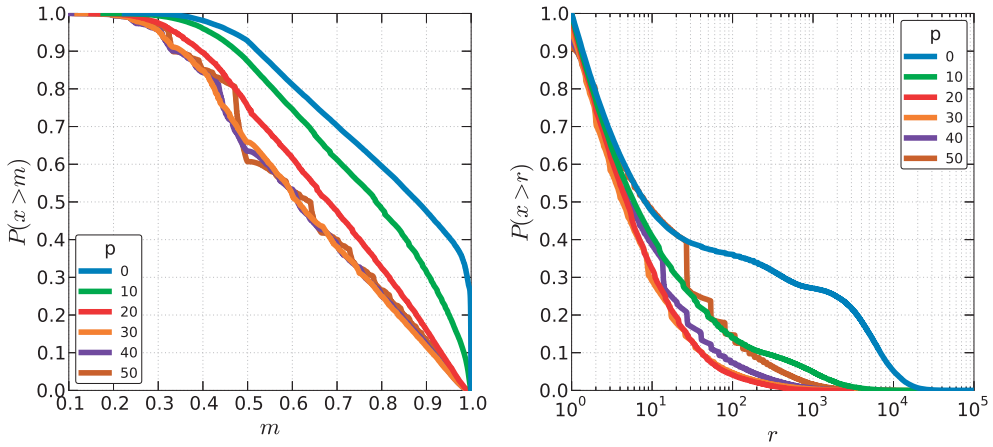

Figure 2.19 – Complementary cumulative distribution functions (ccdfs) of the probability of topics within articles, $\pi(t|d)$, for various percentiles p of filtered concepts. For each article d, we calculate the maximum of the probability $\pi(t|d)$ over the topics , $m(d)$, plotting the ccdf of such values for all the articles in (a). In the same fashion, we calculate the ratio, $r(d)$, between the maximum $m(d)$ and the second highest value of $\pi(t|d)$, whose ccdfs are displayed in (b).

maximum probability m which exceeds 0.5. Hence, the most important topic inside these articles is also dominant. In particular, filtering more concepts decreases the fraction of articles with a dominant topic. This dependence is marked for the first percentiles ($p = 0, 10, 20$) but gets attenuated for the subsequent ones where the ccdfs are closer to each other. Therefore, we may guess that the topic mixture that describes single articles becomes more balanced as the maximum probabilities lower. In order to determine if this is the case, we analyze the ccdfs of the ratio between the maximum and second highest value of $\pi(t|d)$, $r(d)$, calculated for every article d. Figure 2.19 (b) shows that 80% of articles have a ratio r greater than 2 for all the percentiles, meaning that the vast majority of articles are quite focused on the most important topic. Remarkably, even the fraction of articles with $r > 10$ is not so small. Therefore, if we increase the filtering percentile p the topic composition of articles is not fuzzier. However, the behavior of the ccdfs when increasing the percentile p is not the same as in (a). Indeed, the ccdf drops faster when passing from $p = 0$ to $p = 20$, remains very close to the last at $p = 30$ but suddenly raises again for $p = 40, 50$. Such trend of the ccdfs is not monotonic in p, contrary to what noticed in (a). The high jumps in the distributions for $p = 40, 50$ suggest that the values of r

are very concentrated in that regions. From what we have observed so far, it is fairly justified to assign articles to their most probable topic.

The content of the topics for different filtering levels can be analyzed more in detail by means of the Sankey diagram in Figure 2.20. Each topic t, represented as a box, is labeled according to the 10 most frequent concepts within, based on the prominence of words recovered from $\pi(w|t)$. The height of the box denotes the number of articles associated to the topic since it is the most likely for them. Each column in the diagram corresponds to a given level of filtering p for which the most meaningful topics are displayed. A rough comparison of the topics with the communities of article in the Sankey diagram based on community detection, Figure 2.12, reveals that the number of topic is higher than the number of communities at the same percentile. Already when no concepts are removed ($p = 0$) some topics are more specialized than

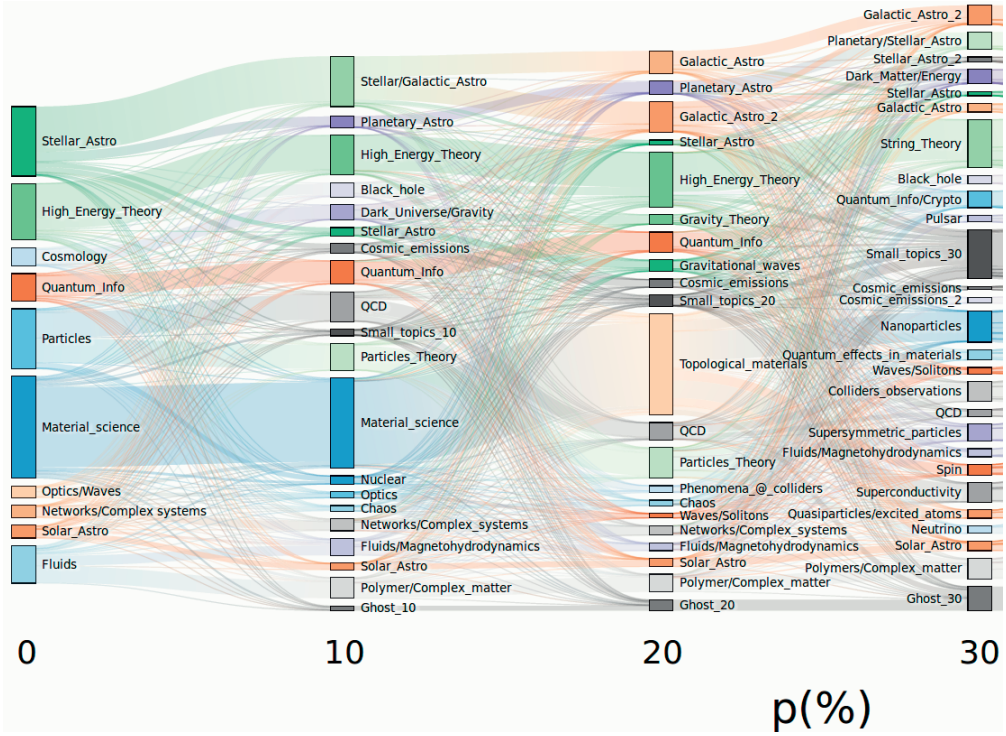

Figure 2.20 – Static Sankey diagram representing the topics found by TM on the physics dataset. Each topic is identified with a colored box whose height corresponds to the number of articles associated to it. Each article d is assigned to the topic with maximum probability $\pi(t|d)$, *i.e.* the topic that describes the highest portion of the article. Topics are manually labeled from the ten most representative concepts according to the probabilities of concepts given the topic, $\pi(w|t)$. For the ease of visualization, only topics with probability $\pi(t) > 0.01$ are shown, whereas the remaining ones are incorporated together in a single "super-topic" denoted as "Small topics". The boxes labeled "ghost" are composed by articles that do not contain any significant concept at a given percentile p, therefore are not part of the dataset used by TM. The thickness of the bands between boxes indicates the number of shared articles. Interactive version available at [236]

the communities of articles; for example, two different branches of astrophysics ("Stellar_Astro"

and “Solar_Astro”) as well as “Cosmology” are present together with other narrow topics like “Optics/Waves” and “Fluids”. However, we also identify topics which are considerably broader, *e.g.* “Material_science”, “High_Energy_Theory” and “Particles”. As p increases, topics become more fragmented: for example, “Particles” evolves into “Particles_Theory”,“QCD” and “Nuclear”, while “Stellar_Astro” splits into “Stellar_Astro”, “Cosmic emissions”, “Stellar/Galactic_Astro” and “Planetary_Astro”. Nevertheless, the fragmentation does not take place for every topic at the same stage as demonstrated, for example, by “Material_science”: the topic remains unaltered up to $p = 30$ for which it breaks down into smaller topics like “Superconductivity” and “Nanoparticles” among the others. The overall phenomenology of filtering resembles the one highlighted in previously for article communities. Interestingly, both approaches shows the presence of a topic/community, “Quantum_Info”, that does not deteriorate by splitting or mixing with others as p grows, meaning that it is somehow isolated since it shares only few concepts with other topics/communities. Intriguingly, this observation may denote the presence of a “cultural hole” between quantum information and the other disciplines. Despite the pruning of generic concepts allows the emergence of finer-grained topics, filtering out too many concepts implies that useful information is ignored. As a consequence of increasing p, topics with small probability ($\pi(t) < 0.01$) grouped in “Small_topics” attract more and more articles.

In order to properly compare the results of TM with those from the article similarity networks, we must remember that the two approaches provide clusters of different type of nodes. Indeed, TM reveals the presence of latent groups of concepts that correspond to topics, while the communities of the similarity networks are composed by articles. In order to compare similar entities, we first investigate the relation between topics and communities of articles focusing on shared concepts. Next, we reverse the viewpoint, examining the articles in communities along with the ones associated to topics.

Concerning the concepts, we characterize each community of articles C by ranking concepts with respect to the frequency among papers. Likewise, concepts within topic t are sorted according to the probability $\pi(w|t)$ which represents the importance of concept w for the description of the topic. For each topic t and community of articles C, we take the concepts in the intersection and correlate their rankings using the Kendall coefficient τ_b, a nonparametric measure of the association between the two rankings [261]. Given the set of K concepts in the intersection, each one is associated to a probability x_k in topic t and a frequency y_k in community C. For any pair of concepts k and l, we say that they are concordant if their ordering is the same in both variables, *i.e.* if $x_k > x_l$ and $y_k > y_l$; or if $x_k < x_l$ and $y_k < y_l$. Conversely, they are discordant if the ordering is reversed, namely either $x_k > x_l$ and $y_k < y_l$; or if $x_k < x_l$ and $y_k > y_l$. Finally, the pair is tied on x (y) if $x_k = x_l$ ($y_k = y_l$). Among the number of possible pairs, $K(K-1)/2$, the number of concordant and discordant pairs are then R and S, respectively, while X_0 (Y_0) denote the number of pairs tied only on x (y). The Kendall correlation coefficient is then defined as

$$\tau_b = \frac{R - S}{\sqrt{(R + S + X_0)\,(R + S + Y_0)}}\,. \tag{2.24}$$

This value varies between -1 and $+1$, where -1 indicates opposite rankings for the two quantities and $+1$ equal rankings. The correlation coefficient τ_b is calculated between every topic and every community of articles after removing the same percentile p of generic concepts. The heatmaps of τ_b are displayed in Figure 2.21. At $p = 0$ single communities are strongly correlated with individual topics as shown by the dark red elements on the diagonal. The same is valid for $p = 10$ even if the correlations are less sharp; curiously, most of correlations outside the diagonal are negative only in this case, albeit their magnitude is quite lower than positive values on the diagonal. The communities about "Particles", "Theory_Q-Gravity" and "Networks/Polymers" have close correlations with two or more topics, corroborating their more specialized nature. At $p = 20$, the coefficients span a wider range of values and most of the communities share a high correlation with multiple topics, again suggesting that part of the topics are more definite. For example, "Astro. (Galaxy/Star)" has 5 topics with close correlations, while for "Particles (Exp.)" the similar topics are 3. Curiously, the community "Graphene" exhibits mild correlations with the topics "Topological_materials", "Polymer/Complex_matter" and "Chaos", while the topic "Topological_materials" is heavily correlated to the community about "Cond_Mat_(Topl_Ins)". Finally, for $p = 30$, we still observe strong correlations of the communities with a small subset of topics whereas the weak correlations that form a kind of background are reduced. For example, the "Astro. (Galaxies)" community shows a high correlation with topics related to astrophysics like "Cosmic_emissions", "Galactic_Astro" and "Stellar_Astro". Therefore, the localized strong correlations demonstrate that there is a tight correspondence between communities and topics in terms of their characteristic concepts, despite the very distinct methods that are adopted to identify these groups. Such overlaps emphasize the presence of an underlying thematic organization of the articles which emerges when both methods are applied.

After we established a relation between the priority of concepts for the communities of articles and topics, we compare here the articles that constitute the communities and those associated to the topics. In particular, articles are assigned to their most important topic as deduced from $m(d) = \max_{t \in T}(\pi(t|d))$. The comparison between articles offers a complementary perspective to identify similarities between communities and topics. The Jaccard score is then used to assess such similarity varying the percentile p of removed concepts, as displayed in Figure 2.22. For $p = 0$, the high Jaccard scores on the diagonal of the heatmap indicate the presence of an almost perfect, one-on-one correspondence between topics and communities. The smaller values on the diagonal are clearly influenced by the disparity between the community and topic size, as indicated in square brackets. For instance, the "Theoretical_Physics" community is almost twice as big as the "High_Energy_Theory" topic. Likewise, the community "Networks" is two times smaller than the topic "Networks/Complex systems" which also includes concepts related to complex systems apart from the ones about networks. The relations highlighted here confirm the correlations between the ranked concepts in the communities and topics. A similar pattern of large scores on the heatmap diagonal is also observed for $p = 10$. Apart from the greatest elements that spot a well defined correspondence between a single community and a given topic, the presence of lower values on the diagonal indicate anyway a good matching of the communities with topics since their values are remarkably higher than the ones in the

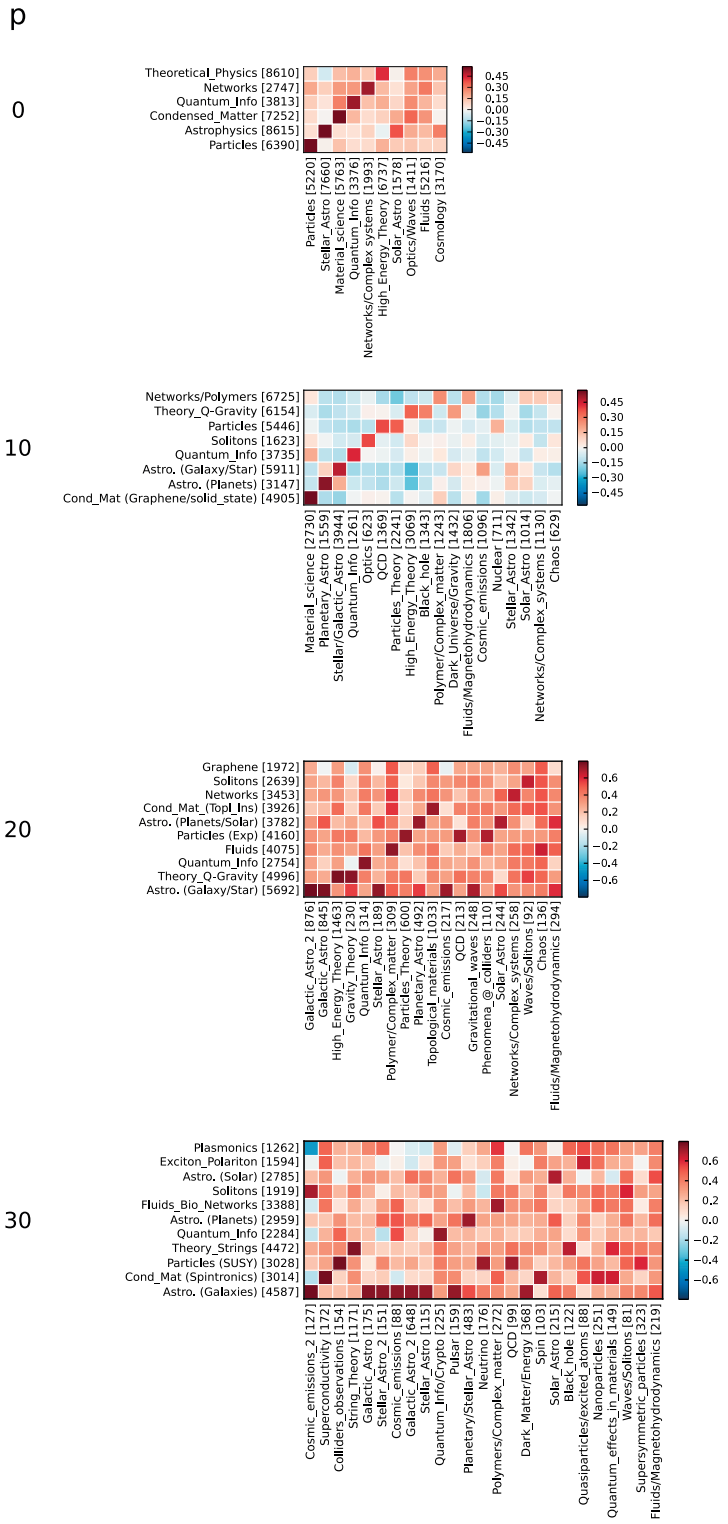

Figure 2.21 – Kendall correlation coefficient τ_b for different topics (columns) and communities of articles (rows) detected at a given filtering percentile p. Only those topics for which $\pi(t) > 0.01$ are included in the heatmaps. The number of concepts in every community or topic is indicated within square brackets. The correlation coefficient is restricted to the intersection between concepts in a topic and the ones within a community. The limits of the color scale for each percentile p are equal in absolute value.

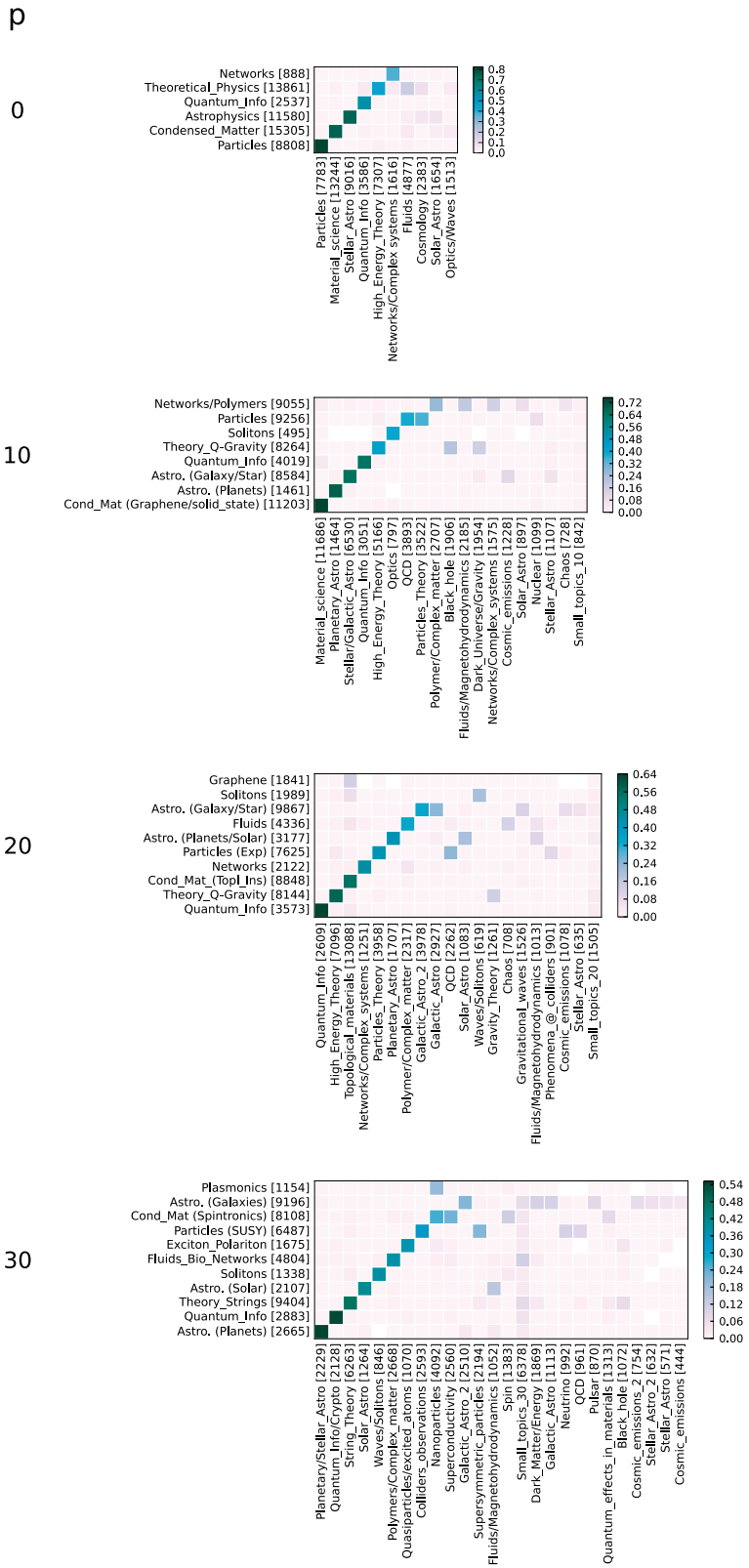

Figure 2.22 – Jaccard score between the articles that compose the communities (rows) and the ones associated to topics (columns) for a given filtering percentile p. Only those topics for which $\pi(t) > 0.01$ are included in the heatmaps. The number of articles in every community or topic is indicated within square brackets.

background outside the diagonal. As an example, the "Theory_Q-Gravity" community overlaps mostly with the "High_Energy_Theory" topic, although it is modestly related to the "Black_hole" and "Dark_Universe/Gravity" too. Moreover, the "Particles" community shares a comparable score with "QCD" and "Particles_Theory" topics. A milder relation is finally present between the "Networks/Polymers" community and the topics "Polymers/Complex_matter", "Fluids/Magnetohydrodynamics" and "Networks/Complex_systems". In this case, the community has a non-negligible overlap with a puzzling topic, "Solar_Astro"; such overlap is due the articles about solar hydrodynamics that are incorporated in the community. For $p = 20$, the communities on the diagonal with a high score, in the lower left part, are less than the previous percentile. The other communities with a moderate value of the score on the diagonal (colored in blue) display a bit higher numbers of topics with a mild overlap off the diagonal. The "Solitons" community has a maximum score which is pretty weak but all the other scores are rather lower than it, the same happening with the "Graphene" community. Finally, in the case of $p = 30$, we notice, again, that every community has an higher score with a distinct topic which allows to delineate a correspondence between them: albeit the maximum score for a community may not be as big as the maximum of the whole heatmap, it is dominant with respect to the other scores in the same row. In this case, an exception to this trend is given by the community "Cond_Mat (Spintronics)" which has a similar overlap with the topics "Nanoparticles" and "Superconductivity", plus smaller overlaps with "Spin" and "Quantum_effects_in_materials". Other two exceptions are the communities "Astro (Galaxies)" and "Particles (SUSY)".

Overall, the detailed comparison of the Jaccard scores reveals that there is a strict relationship between every community and at least one topic in terms of common articles, as highlighted by the strong similarities on the diagonal of the heatmaps. Indeed, these findings allow to draw a one-to-one correspondence between communities and topics in most of the cases. Nonetheless, the presence of such correspondence is not trivial since the communities of articles and the topics have been unveiled from very different methods based on clustering articles and concepts, respectively. In addition, the assignment of articles to topics has been devised in a strict way associating each article to its most important topic, thus overlooking the information about the topic mixture that describe the content of the article. Despite this rough approximation introduces a bias in the assignments, the resulting Jaccard scores disclose a remarkable affinity between communities and topics. These results match closely the high correlations discovered between the ranked concepts in communities of articles and topics. To summarize, both the correlations (between concepts) and the Jaccard scores (between articles) for different communities and topics confirm the existence of a latent organization in peculiar themes within physics. Interestingly, the results from the two methods are coherent in the identification of the more pronounced relationships even if the methods are grounded on distinct criteria.

After comparing the set of articles that compose the communities against the articles associated to topics, we can ask if these groups of articles match some known classification. The answer to this question is important since it eventually provides an independent way to validate the discovered

groups of articles[10].

2.3.4 Comparison with the ground-truth

The articles deposited on `arXiv` are classified by the authors with a primary category (mandatory) and secondary category(ies) (optional). In our case, articles have been selected from primary categories that correspond to physics subject classes, as detailed in Table 2.1 and Figure 2.1. Each subject class allows to define a ground-truth for each article that corresponds to its primary category. For every community of articles or topic we then evaluate its correspondence with the ground-truth categories. However, due to historical reasons related to the popularity of `arXiv` in different fields of physics, they are not homogeneous in the number of papers they contain and the categories do not necessary reflect a well-principled division of the papers into hierarchically similar areas. Therefore, we decided to group together some categories (like the high energy physics) into macrocategories that correspond to a more homogeneous classification of the different fields. Nonetheless, we also consider a finer-grained division of the articles subcategories that in the nomenclature of `arXiv` are specified by a dot after the category name; *e.g.* `astro-ph.CO` is the subcategory about cosmology and non-galactic astrophysics within the astrophysics (`astro-ph`) category. Once we clarified the ground-truth classification in use, we analyze the matching pattern between groups of articles and categories by means of the Jaccard score between them, as defined in Equation 1.20. The extensive comparison between communities and categories is outlined in Figure A.3. Here, however, we are interested in analyzing two different measures that are derived from the Jaccard score: the *recall* and *precision* scores. The first quantifies to which extent a known category is well reproduced by a discovered group of articles; it is therefore defined for every known category C_i as the maximum Jaccard score of the category over all the discovered groups D_j:

$$R(C_i) = \max_{D_j} J(C_i, D_j)\,. \tag{2.25}$$

The score is close to one if a discovered group matches fairly well the known category. Taking the specular point of view we then define the other quantity, the *precision score*, which determines if a discovered group is well represented by a known category. Such score is indeed defined as the maximum Jaccard score between the discovered group D_j over all the known categories:

$$P(D_j) = \max_{C_i} J(C_i, D_j)\,. \tag{2.26}$$

Then, we calculate for every know category and every discovered group the respective score and we evaluate the overall quality for each score by ranking it in increasing order and displaying

[10] Unfortunately, such comparison is possible only for the articles in `arXiv` and cannot be broaden to the concepts since a categorization is missing. A very interesting and promising effort in this direction is actually taking place for the American Physical Society journals where the authors that publish a paper are asked to label it using the so-called *PhySH* (Physics Subject Heading) in order to describe the content of the paper with a wide range of features such as the techniques, the methods, the physical system that they analyze and so on. [262]

the respective value. The rankings are normalized by the number of known categories in the case of the recall and by the number of discovered groups in the case of the precision. The trend of the scores is then analyzed varying the percentile p of removed concepts in order to understand the effect of the filtering on the two scores. Figure 2.23 shows the resulting scores. In the case of macrocategories, the recall score has very close values in the left part of the plot, *i.e.* for the lower positions. An exception to this trend is observed for $p = 10$, whose values are higher both for the communities of articles and the topics. In the right part, the recall reaches different values when varying the percentile and, in general, the scores for the communities are greater than the ones for the topics in TM at the same percentile. In this interval, however, a clear trend is present: increasing the percentile p reduces the value of the recall. For the subcategories the recall starts with very low values, likely due to the small number of papers that compose them. Curiously, the recall is slightly greater for the articles associated to the topics and does not decrease with p. Indeed, it attains the highest values over the whole rankings for $p = 10$ followed closely by $p = 20$. Then, for $p = 30$, the values lower again as for $p = 0$ and gets closer to each other for the topics and communities. Nevertheless, the variation of the scores for different percentiles is much less pronounced than in the case of macrocategories. The precision score for the macrocategories exhibits a more clear trend when p raises since the curves are more separated and the precision values are reduced. The results from TM are again lower than the ones from the communities of articles. For the subcategories the precision values show again a decreasing

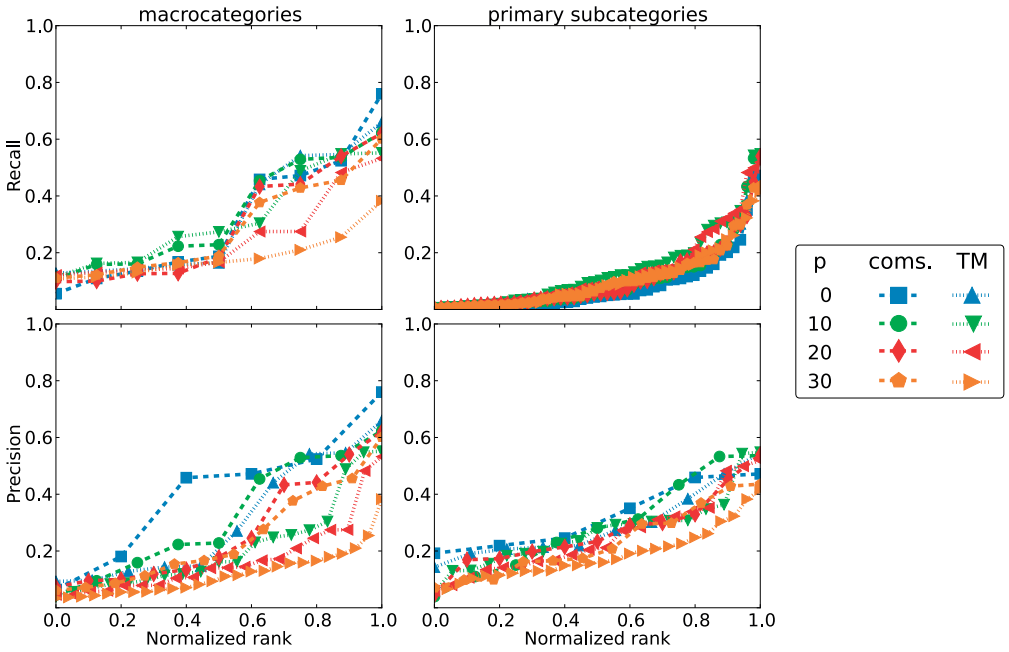

Figure 2.23 – Ranked values of the recall and precision scores for the ground-truth classification of articles in macrocategories and subcategories of `arXiv` . The color denotes the percentile p of removed concepts. The dashed lines refers to the communities of articles and the dotted lines to the articles associated to topics by TM.

behavior as the percentile p increases, but the values for the communities of articles and the

topics are not well separate as before. The overall values of the recall and precision scores suggest that the communities of articles have a better performance in most of the cases. Nonetheless, the scores are not tremendously higher than in the case of articles assigned to topics. These findings, however, are not completely unexpected: as several studies highlighted, the problem of identifying known communities only relying on the structural properties of a real network seems not to have acceptable solutions [163]. The cases where the performance of the algorithms is good are limited to synthetic networks with a planted community structure. In most of the real networks, integrating the information about metadata like the ground-truth above is then necessary to achieve a good performance, at least in some scenarios [263].

Nevertheless, it is important to stress a crucial difference, very often overlooked: metadata information of real networks should not be considered having the same reliability of the ground-truth in synthetic networks. Indeed, a ground-truth is enforced in the structure of synthetic networks by construction while metadata may not be the underlying basis from which communities formed [264]. On the other hand, metadata are usually termed ground-truth and wrongly treated as such. However, metadata may only help to achieve a better community structure, although there is no guarantee that this is case. Accordingly, we cannot become obsessed by reproducing the metadata in the community structure although their interplay with the communities is, indeed, a valuable contribution. Ergo, the quality of the community structure must not be evaluated solely in terms of metadata but other criteria should be considered when possible.

In the above case, for example, examining the concept frequency within communities (as reported in the Sankey diagrams [236]) highlights that coherent topics are identified in the communities. However, the fact that the communities do not match perfectly the categories is not surprising. Exploiting the concepts in the full text of the articles – without restricting to title and abstract – allows to delineate more in depth the content, catching the complete spectrum of background knowledge, methods, techniques, and applications of the research. Some of these concepts can be shared among different fields that do not fall within the same category, especially if the research is interdisciplinary or concepts borrowed from one field are used in another one. A remarkable example of this case is the recent developments of AdS/CFT correspondence from string theory in condensed matter physics [265, 266]. The comparison of the communities of articles with the state-of-art technique in topic modeling, *i.e.* TopicMapping, suggests a certain robustness of the detected structure, given that similar topics are also discovered from such complementary approach. Therefore, we may argue that there is a core organization structure unveiled by both methods. Perhaps a closer matching of the communities with existing categories would have been reached, for example, if we had limited the semantic similarity to title and abstract, since they are less rich in terms of concepts or establishing article connections from citations. Notwithstanding, simply reproducing the categorization of articles already known would be of limited interest. In conclusion, we think that it is more enlightening to uncover unexpected relationships that go beyond the available information, an achievement that has been possible thanks to the method for filtering generic concepts.

2.4 Conclusions and remarks

The availability of large collections of documents offers the opportunity to investigate the semantic information presents within them. From this large-scale analysis is then possible to characterize the complex pattern of word adoption in texts, not only examining where they are used (the frequency among documents) but also how (the term frequency inside documents). This analysis is important to identify documents with similar content in order to automatically classify them in thematically related groups. In this way, we can explore the emerging meso-scale organization of documents in different topics.

In the present Chapter we focused on a corpus of scientific preprints where we discovered that concepts extracted from their body are not equally valuable to describe the content of the articles. Indeed, *generic concepts* play the role of *buzzwords* that do not carry a specific meaning but are quite vague and indefinite, being present in various articles regardless the topic that they address [267, 268]. Their presence dramatically hinder the construction of genuine similarities between articles since they are responsible for spurious contributions to the link weights. For that reason, the resulting density of the article similarity network is very high, as detailed in section 2.1. As a consequence, the community structure is composed by major topics drawn together by common concepts which do not allow to recover finer-grained communities.

Contrary to what expected, the notion of generality of a concept is not naively related to the frequency among articles but it depends on a quantity that captures more deeply the information about the distribution of a concept, *i.e.* the entropy. Guided by the empirical evidence that concepts tagged as *common* possess an higher entropy, we have developed a method that accounts for such trend in order to quantify the generality of concepts. Firmly grounded in the maximum entropy principle, the proposed method allows the identification of generic concepts based on the difference between the maximum entropy and the actual one, as described in section 2.2. The advantages of this method are at least two: first, it is unsupervised as it does not need the validation of experts in order to (manually) spot generic concepts. Second, it provides a measure of generality that is intrinsically related to the context since it depends on the corpus under scrutiny. As an example, the method can be applied recursively to sub-corpora in order to identify generic concepts within a given topic. Removing a fraction p of generic (uninformative) concepts, the noise captured in the article similarities is reduced since only *relevant concepts* contribute to the link weights. In this way, the article similarity network becomes sparser causing the detected communities to represent more closely specific topics which offer a detailed overview of the thematic organization of large corpora, as reported in section 2.3. Filtering generic concepts, however, is useful also in the case of topic modeling, a complementary perspective to determine the topics in a corpus. In such case, the granularity of topics (represented by groups of words) also improves as a direct effect of the removal of generic concepts. Given the division of articles in various categories within `arXiv` , the emergence of a thematic organization is expected, although neither the communities of articles nor the modeled topics match very closely the categorization.

The validity of the filtering method is not limited to the case of a curated ontology of crowdsourced

concepts but it applies also to the removal of irrelevant keywords extracted from webpages about climate change, as outlined in subsection A.2.2. Regardless the differences in the structure, composition, and purpose of the kind of documents, the method effectively improves the topic description of the latter, even if the results are less outstanding due to the rough nature of keywords as compared to scientific concepts.

As a general purpose method, the characterization of keywords by different degrees of generality might be helpful in order to propose a sorted pool of words when building an ontology for a corpus of documents or to improve an existing one. Among the efforts to characterize more precisely the vast knowledge enclosed in scientific publications, a recent development has been introduced by the American Physical Society with the Physics Subject Headings (*PhySH*) system, a hierarchical annotation framework that superseded the old PACS with the aim of classifying the manuscripts by the various facets that a publication addresses, from methods and techniques that are adopted to the physical systems that are studied and the research areas that are concerned [262].

Regarding scientific concepts, an interesting question concerns the evolution over time of the concept generality in order to establish if some trend is present, *e.g.* concepts that were specific and then gained momentum or, viceversa, concepts that were abandoned after being mainstream for a period of time. In the first case, for example, we can include "graphene" as it has become an increasingly popular material, very promising for many technological applications. In this perspective, we devote the next Chapter to the analysis of the temporal trajectory of concepts.

3 Temporal evolution of scientific concepts

Understanding the changes in scientific knowledge without considering its evolution across time is impossible. Indeed, the changes that take place in the scientific research over time are a valuable asset to describe the advances and the specialization of the scientific knowledge. In the literature, several works address this aspect considering the concurrent transformations of paper and author networks [40], the variation of the citations received by papers [52, 54, 55], the transitions of topological indicators in the networks of citations among scientific publications to discover emerging fields [34, 269], and the development of collaboration networks to outline the history of scientific domains [22, 35]. Notwithstanding, none of these studies leverages the semantic content of the articles. In principle, this operation can be realized by considering the evolution of scientific concepts in order to investigate another facet of the progress of scientific knowledge. Thus, an interesting question is to characterize how the consumption of scientific concepts evolves over time focusing on their term-frequency distribution within papers. The method introduced in the previous Chapter exploits precisely such distribution by linking the generality of a concept with its residual entropy, adapting to the corpus of articles under scrutiny. Given the fact that concepts are not exploited always in the same way over time, we guess that their generality across time will fluctuate as well. Accordingly, the temporal traces of concept generality can be employed to outline common trends in their evolution.

To this purpose, we decided to focus on one of the oldest categories of `arXiv` , astrophysics (`astro-ph`), which has been actively used to submit preprints since the birth of `arXiv` in 1992 [270, 271]. Therefore, the long history and the high submission rate per year guarantee that we have good statistics in order to evaluate the concept generality with confidence. The number of submitted papers and the number of unique concepts per year are displayed in Figure 3.1. We consider a 20 years period in time spanning from 1994 up to 2014, since the number of articles in 1992 and 1993 does not provide a considerable coverage, being 59 and 488 respectively. Furthermore, we keep only those concepts appearing in more than one paper and whose term-frequency distribution is different from a delta. In this way, only those concepts with an observed entropy greater than zero are selected to be further analyzed. Every year is studied as a separate corpus containing a given number of selected concepts for which we compute the residual entropy

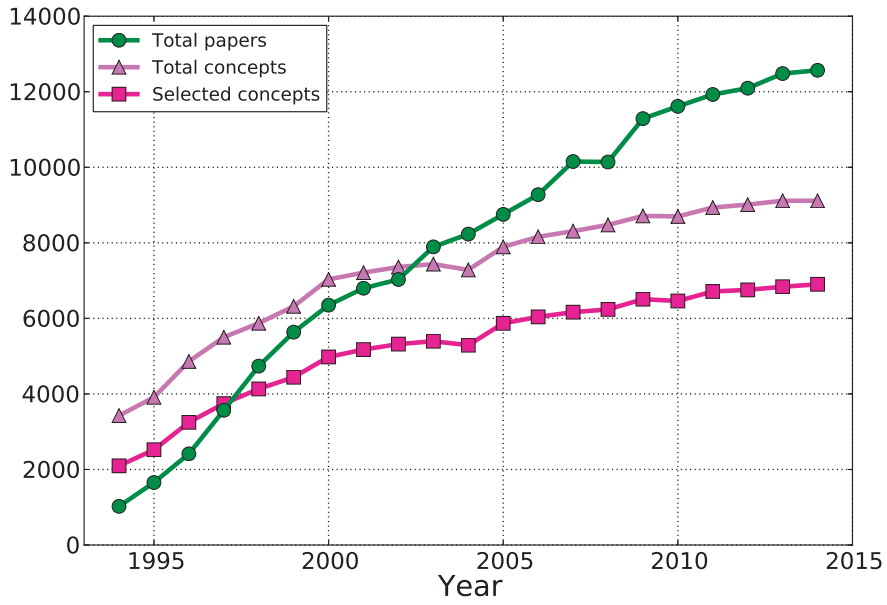

Figure 3.1 – Evolution of the number of articles (○), concepts (△) and selected concepts (□) across the years.

as described in the previous Chapter. Since the raw value of the residual entropy, S_d, cannot be used to compare the "generality" across time, we compute its percentile. We consider percentiles going from the 10th up to the 100th in steps of 10 and assign concepts to their highest percentile. Loosely speaking, we refer to the resulting groups as percentile slices since they contain concepts that are included in a given percentile but not in the previous one. This provides a rough separation of concepts into classes of generality, where each class contains the same number of concepts.

Following the variation of the percentile slice through time, we can characterize the changes of the concept generality. The percentile slices and maximum entropy parameters for selected concepts are displayed in Figure 3.2. We note that the variations of the maximum entropy parameters, s and λ, seem erratic but they may compensate in the maximum entropy of Equation 2.16, implying that the percentile slices p remain constant in some cases. More in detail, we inspect the trend of the percentile slice over time for some reference concepts which are pertinent to astrophysics, providing a comparison with their generality for the physics corpus analyzed in the previous Chapter. In particular, concepts tagged as common by SW experts, like "Energy" and "Universe", always fall into the 10th percentile from 1996 onward. This observation suggests that they can be considered as very generic at any time. Moreover, they are also included in the 10th percentile for the physics corpus examined in the previous Chapter, being general both for the big corpus of physics articles and one of its subcorpus, astrophysics. Nevertheless, the concept "Star" is considered very generic for most of the years and also in the physics corpus, albeit it has not been tagged as common. Again, the method based on the residual entropy is able to spot automatically those concepts that are generic for a given corpus. Important concepts more related to astrophysics, from "Hertzsprung-Russell diagram" down to "Perfect fluid" are also outlined.

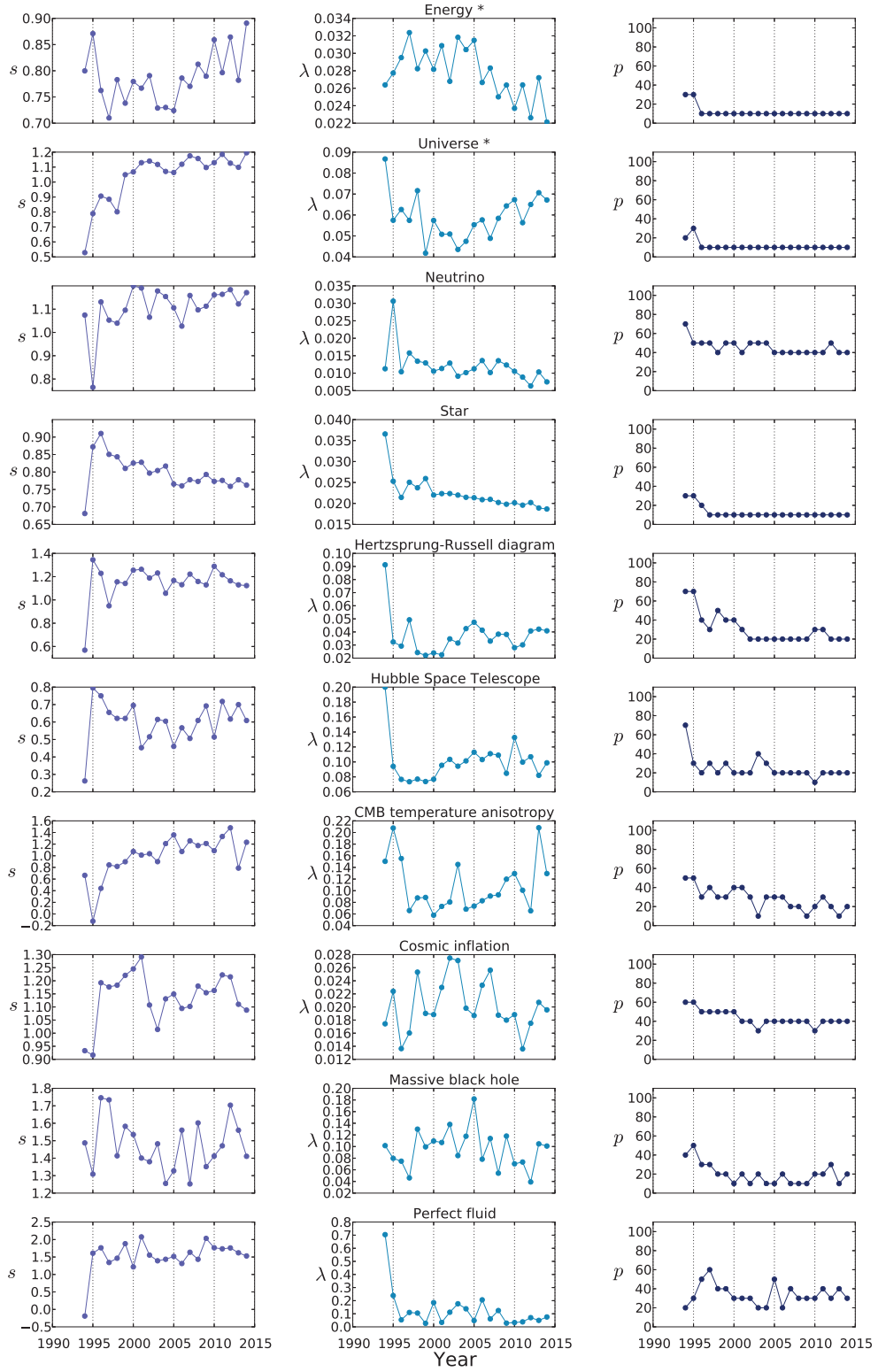

Figure 3.2 – Variation over the years of the maximum entropy parameters and the percentile slices of selected concepts. Every row represents a different concept; those with an asterisk are the common ones as tagged by SW experts. From left to right, the columns show the maximum entropy parameter s associated to the power-law, the parameter λ responsible for the exponential cutoff, and the percentile slice evolution, p.

The temporal traces of the percentile slice fluctuates more in these cases. Another concept with a prominent role in astrophysics is "Dark matter". This type of matter is called dark since it does not emit light, therefore is not possible to investigate it by means of standard instruments like telescopes adopted in astrophysics. Yet, the standard model of cosmology indicates that it makes up about 27% of the Universe [272]. The investigation of such mysterious matter has been increasing over the years thanks to the availability of advanced instrumentations like the Planck telescope, among others . Therefore, many models have been proposed to justify the presence of dark matter in the Universe [273–276]. Such models postulate the existence of different types of dark matter characterized by various properties and each type can be associated to a characteristic concept. The evolution of the percentile slice for some of these concepts is shown in Figure 3.3. Interestingly, "Dark matter" alone is a concept always generic and a similar behavior can be observed for "Cold dark matter". On the other hand, "Warm dark matter" and "Weakly interacting massive particle" may be regarded as very specific over the whole time. Finally, other flavors of dark matter exhibits very different trends, some more constant while others more oscillating.

Since the traces of the percentile slice evolve disparately, we want to understand if it is possible to identify characteristic trends in such traces that may describe concordant evolutions of a considerable number of concepts. For example, few plausible trends of concept generality are sketched in Figure 3.4. Notwithstanding the intuitive description of these trends, there is no reason why concepts should follow them. In principle, the trends that mostly represent the percentile slices could be completely different. Therefore, to avoid imposing any particular trend, we need to resort on a method that automatically determine the most recurrent patterns. Such problem can be though in terms of finding few typical patterns that explain common trends observed in empirical data. Dimensionality reduction techniques are a large class of methods that accomplish this task [277, 278]. To explain the basic principles, let us consider a matrix $\boldsymbol{X}$ of dimension $C \times T$, where C is the number of concepts and T is the number of years. In the following, we focus only on those that are present for every year in the time interval under study. The element x_{ij} is the percentile slice of concept i in year j. Any dimensionality reduction scheme tries to approximate the matrix $\boldsymbol{X}$ with a product of two or more matrices that offer a reduced description of $\boldsymbol{X}$. Ideally, the approximation allows to reconstruct the entries in the original matrix $\boldsymbol{X}$ by capturing relevant components of the observed variations. One popular method applied for time series analysis is Singular Value Decomposition (SVD [244], see Equation 2.21) which aims to decompose the original matrix as follows:

$$\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T = \sum_{k=1}^{K} \lambda_k \boldsymbol{u}_k \cdot \boldsymbol{v}_k^T . \tag{3.1}$$

Each of the K components of $\boldsymbol{V}$, $\boldsymbol{v}_k$, identifies a direction of optimal projection in a low, K dimensional space along which the values in $\boldsymbol{X}$ tend to align. Moreover, the eigenvalues λ_k quantify the contribution of the respective component k to the approximated description of $\boldsymbol{X}$. In our case, the columns of $\boldsymbol{v}_k$ provide the pattern of variation most frequently observed that reconstruct the traces in $\boldsymbol{X}$ more faithfully. However, none of the extracted components describe a prototypical series that is somehow regular. Indeed, each of them basically encodes the highly

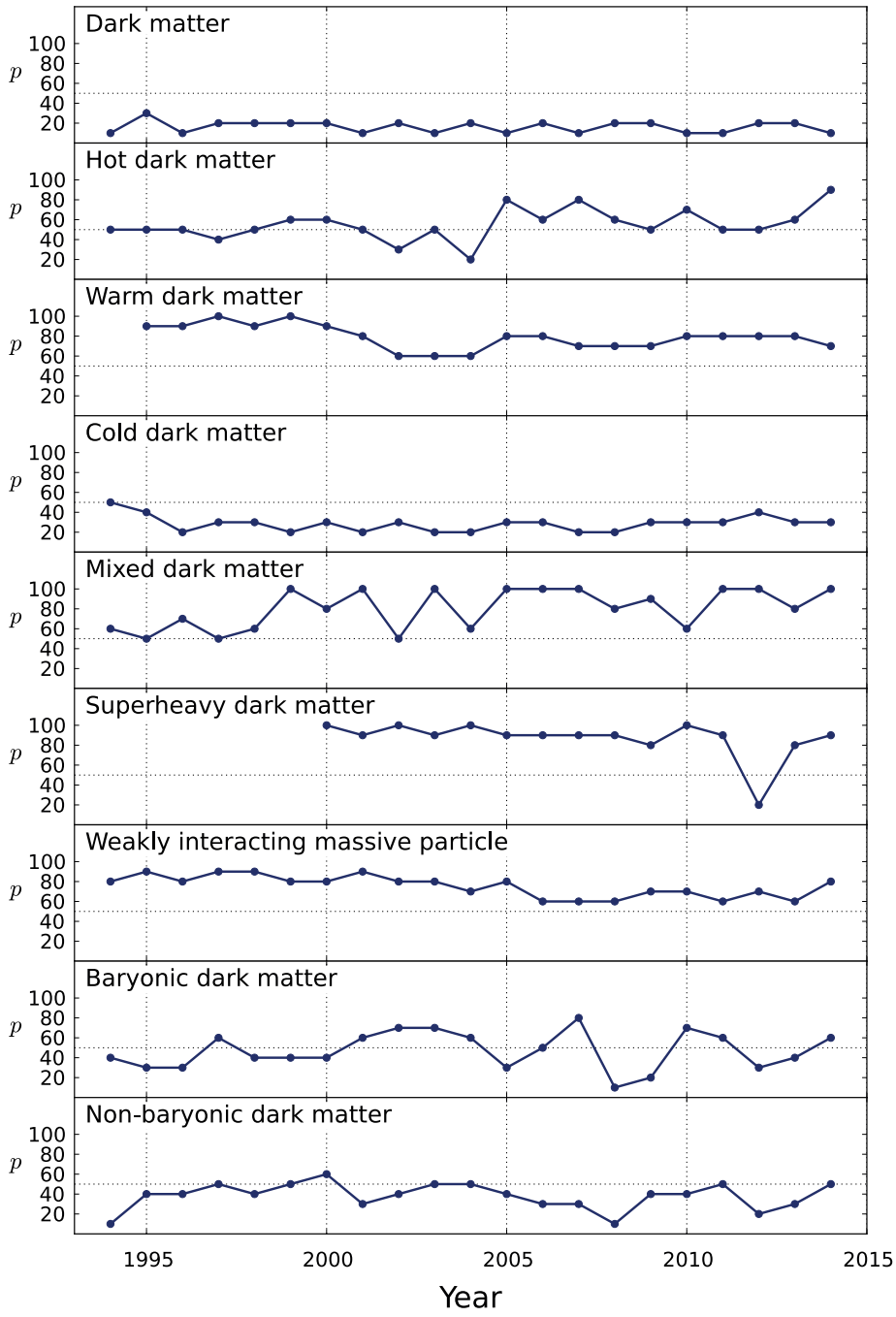

Figure 3.3 – Evolution of the percentile slice, p, across time for different concepts associated with "dark matter".

fluctuating nature of the percentile slice that most concepts exhibits. Hence, the characteristic patterns do not show interesting trends over time but simply capture the noisy variations of concept generality. Nevertheless, we also tested a more advanced technique in order to establish if common trends in the percentile slice evolution can provide a classification of concepts. Such

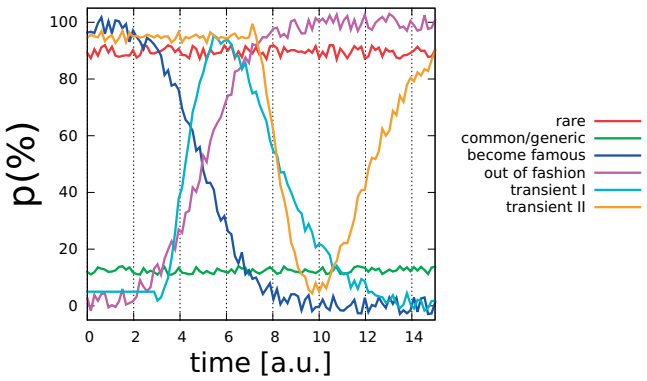

Figure 3.4 – Schematic portrait of plausible trends in the evolution of the percentile slice of concepts. A brief description of the different traces is provided in the legend.

technique, called Dynamic Time Warping (DWT, [279]), has been developed specifically for time series analysis, aiming to compute a distance between time series and then clustering together those that are closer to each other. Unfortunately, also this attempt did not give meaningful results. We claim that these analysis were not successful since the traces of the percentile slice are very short in time – only 20 points – whereas the techniques adopted in time series usually deal with much longer traces. Moreover, fluctuations in the time series are usually smooth with a small amplitude, while in our case they are oscillating up and down many times with big jumps.

In conclusion, despite the identification of characteristic variations of concept generality has not been accomplished, we have been able to explore more closely the role of few important concepts in astrophysics that remain general over the years or, on the contrary, are quite specific throughout the whole 20 year interval. An interesting direction of further investigation would be to analyze more carefully the role of the power-law exponent derived from the maximum entropy principle, which shows a typical peak around $s = 3/2$ for the concepts in the physics corpus (see Figure 2.7). This remarkable maximum is especially compelling since it is the exponent typical of critical branching processes à la Galton-Watson [280]. In this framework, it is intriguing to imagine that old articles "inspire" (or "generate", in the language of branching processes) new manuscripts that inherit roughly the same number of concepts and similar term-frequencies from the old ones. For such kind of analysis, the temporal framework described above is the ideal one. Testing if a critical branching process is responsible for the content of new articles would be interesting since it may shed light onto the mechanism that drives the evolution of scientific knowledge.

Conclusions and outlooks

Investigating the scientific knowledge as a large-scale phenomenon requires the contribution of scholars with very different backgrounds, since every discipline involved (from sociology to mathematical modeling) provides its own viewpoint on the study of science of science. In particular, the unprecedented disposal of digital publication data paves the way to investigate the scientific knowledge from a quantitative perspective, leveraging multiple types of information contained in articles. Many studies consider the citation pattern among publications as the key ingredient to describe, for example, the organization of science in domains. However, few of them actually consider the semantic content of articles which outlines the research work therein.

In this Thesis, we offered two different contributions to characterize the scientific knowledge from the semantic content of articles. In particular, we analyzed the relationships between the concepts present inside the full text of articles and the articles themselves as a complex network. Considering the content similarity between articles, we showed that the resulting network may be extremely dense, obfuscating the underlying organization of articles in thematically related groups. Such high amount of interactions stems from the presence of the so-called buzzwords, *i.e.* terms that are not very relevant to describe the content of scientific articles since they are pervasively adopted in many articles from various fields. These generic concepts contribute to create spurious similarities between articles that would not be so much related otherwise.

Our original contribution amounted to understand how the microscopic usage of scientific concepts inside papers is related to the "generality" of concepts through their residual entropy, a rigorous measure grounded in information theory that quantifies the intrinsic semantic information encoded in scientific concepts. Therefore, instead of reducing the number of concepts using naive – but intuitive – methods, like considering only the most frequent concepts inside a document, we deepen the comprehension of the relevance of scientific concepts in order to remove the less informative ones. The filtering of concepts enhances the similarity relationships, thereby the community structure of the similarity network portraits more specific "subjects". The same filtering approach is valid when characterizing the composition of articles in terms of latent topics, *i.e.* groups of concepts, that describe the article content. Therefore, combining the macroscopic description of article and concept similarity networks with the microscopic analysis of the concept relevance provided an original contribution to the understanding of the emerging organization of scientific knowledge. These findings are particularly relevant when studying large corpora

of scientific documents that are not divided in predetermined categories or lack a fine-grained division, since we can potentially uncover the thematic organization of these documents. This achievement is possible because the "generality" of a concept is not carved in stone but it adapts to the corpus under scrutiny. As a consequence, such concepts are identified more efficiently than resorting to the judgment of experts. Indeed, we proved the presence of false positive concepts that have been marked as common by experts although they are not, as well as false negative, *i.e.* concepts that experts have missed out to tag as common. Moreover, the entropy-based measure used to quantify the relevance of concepts does not rely on the division of a text in parts, therefore it can be applied to unstructured texts like web pages as well, even in the absence of an ontology of concepts.

A further contribution to characterize of scientific knowledge was provided by considering its temporal evolution. Indeed, the advancement of science manifests with the emergence of new scientific paradigms and discoveries that modify the knowledge landscape. In this global picture, the evolution of the consumption of scientific concepts in terms of how they are employed shed light into the historical development of science. In this case, quantitative studies of the novelties emerging in science would need to be supported with qualitative studies touching the sociology, history, and philosophy of science. Indeed, the direct human knowledge is fundamental to track historical trends in science since a complete understanding of the mechanisms that govern its evolution as a complex system is far from being reached.

An interesting direction of further investigation would be to understand the relation between the paper similarity networks after filtering generic concepts and the map of scientific papers defined in terms of cultural holes [211], since the latter approach is also based on an entropic measure to quantify the dissimilarity of scientific jargon. Regarding the scientific concepts, we can try model their interaction as an Ising-like spin system where concepts are present or not in articles, looking at the possible couplings between concepts that describe the strength of their interactions [281]. Moreover, the same principle to characterize generic concepts can be applied to other ontologies like the *MeSH* (Medical Subject Heading), a well-curated classification system of medical terms from the U.S. National Library of Medicine, in order to characterize the usage of medical vocabulary [282].

The interplay between scientific concepts and articles can be studied more in depth, looking if a nested structure is hidden in their bipartite network [102, 283]. Such pattern of interactions may highlight the presence of widespread concepts used in papers from diverse fields, as well as specific concepts adopted in particular ones, possibly discovery a relation with generic concepts defined from the residual entropy.

More in general, if we look at the scientific knowledge as a complex evolving system with many types of interactions (between authors, publications, institutions), it is very natural to imagine that our understanding of such system would benefit from taking into account these various interactions simultaneously, since a complex system usually exhibits non-trivial features that can be explained satisfactorily only when taking advantage of the whole complexity. For example, we

could investigate the interplay between the semantic similarity and citation network, discovering communities of different articles based the two criteria and comparing their traits. Another example would be to delineate the scientific concepts that authors use and how they change over time, thereby describing the evolution of the scientific interests of scholars at different stages of their career.

A Entropic selection of concepts in networks of similarity between articles

The majority of the work contained in this Appendix is included in [222, 223]

A.1 Theory

In this Section we provide the analytical details behind the filtering method based on the maximum entropy principle. We begin proving the relation between the full entropy, S_f, and the conditional one, S_c, and then we motivate the decision to use the latter for the design of the filtering methodology (subsection A.1.1). After that, we provide the details of the maximum entropy models introduced in chapter 2 (subsection A.1.2), and we show the equivalence between the residual entropy, S_d, and the Kullback-Leibler divergence, $D_{\mathrm{KL}}(p||q)$, between the probability distribution of empirical observations, $p(k)$, and maximum entropy model, $q(k)$ (subsection A.1.3). Finally, we present the comparison between the concept lists ranked according to residual entropy S_d and IDF (subsubsection A.1.4.1).

A.1.1 Relation between full entropy and conditional entropy

The probabilistic formulation of the entropy adopted in section 2.2, does not contemplate as a possible event the *absence* of a concept c in a document α (*i.e.* $tf_c(\alpha) = 0$). For this reason, in Equation 2.5 the sum starts from $tf_c(\alpha) = 1$. We labeled such entropy, S_c, as *conditional* since it is computed under the condition that concept c appears in the document, thus implying an associated probability $p_c(k) = \frac{N_c(k)}{N_c}$. Nevertheless, we can also define another probability distribution that incorporates the absence event, which translates into the so-called *full entropy* S_f. To construct such distribution, we consider the total number of papers in the corpus, N_a, while concept c appears only in $N_c \leq N_a$ papers. Then, we extend the tf_c probability distribution by including the absence event as a term that corresponds to the fraction of papers where the concept c did not appear, $1 - df$, where $df = \frac{N_c}{N_a}$ is nothing else than the document frequency of concept

c. In conclusion, the probability that the concept c appears $k \in [0,\infty]$ times is $p_f(k) = \frac{N_c(k)}{N_a}$, where $p_f(0) = 1 - df$. As a result, the full entropy associated to distribution $p_f(k)$ is:

$$S_f = -\sum_{k=0}^{\infty} p_f(k) \ln p_f(k) =$$
$$= -p_f(0) \ln p_f(0) - \sum_{k=1}^{\infty} p_f(k) \ln p_f(k) = -(1-df)\ln(1-df) - \sum_{k=1}^{\infty} \frac{N_c(k)}{N_a} \ln\left(\frac{N_c(k)}{N_a}\right).$$

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

$$\begin{aligned} S_f &= -(1-df)\ln(1-df) - \sum_{k=1}^{\infty} \frac{N_c(k)}{N_c} df \ln\left(\frac{N_c(k)}{N_c} df\right) = \\ &= -(1-df)\ln(1-df) - df\ln(df)\sum_{k=1}^{\infty}\frac{N_c(k)}{N_c} - df\sum_{k=1}^{\infty}\frac{N_c(k)}{N_c}\ln\left(\frac{N_c(k)}{N_c}\right) = \\ &= S_{bin} + df\, S_c\,. \end{aligned} \tag{A.1}$$

where we used the normalization condition over $N_c(k)$, *i.e.* $\sum_{k=1}\frac{N_c(k)}{N_c} = 1$. The full entropy, S_f, is then 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 corpus. The second is the entropy computed in Equation 2.5 of section 2.2 modulated by the factor df.

At this point, it is natural to ask whether S_f could be used to classify concepts or not. To this aim, in Figure A.1 we outline the relation between S_f and several quantities in order to establish if it can be adopted as a valid alternative to S_c in discriminating generic concepts. The analysis of the influence between S_f and S_c (panel A), $df \cdot S_c$ (panel B) and $\frac{df \cdot S_c}{S_f}$ (panel C) reveals that a clear separation between common concepts (in black) and the others (in orange) is not present. Performing the same investigation for the first term in Equation A.1, S_{bin}, outlines that S_f does not display a characteristic dependence for the common concepts neither on S_{bin} (panel D), nor on its fraction explained by S_{bin}, $\frac{S_{bin}}{S_f}$ (panel G). Summarizing, none of the dependencies shown in Figure A.1 seems to provide additional solutions to conceive a classification scheme. Ergo, the full entropy S_f is inadequate to distinguish generic concepts, since its discriminative power is not as strong as one of the conditional entropy.

A.1.2 Maximum entropy models

Information theory provides a framework to characterize in a rigorous way the information content of concepts, as encoded by the conditional entropy S_c. However, such raw value, which measures the actual information present in the data, is not sufficient to determine if a concept is generic or not. Indeed, we need to fairly compare the observed entropy, S_c, to an expected value in order to establish if it is small or not. This theoretical counterpart of the observed entropy is the maximum entropy associated to an expected probability distribution where some of its

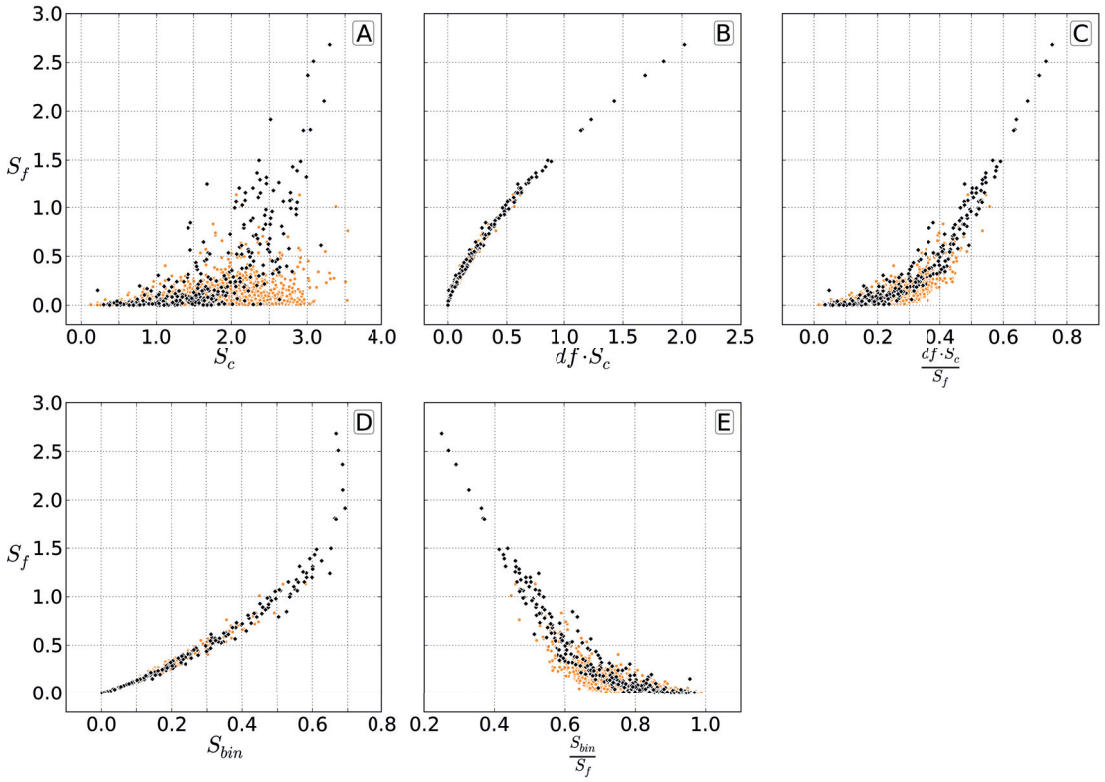

Figure A.1 – Relation between the full entropy S_f and the conditional entropy S_c (A), the contribution of the conditional entropy to the full one $df \cdot S_c$ (B), the conditional entropy contribution over the full one $\frac{df \cdot S_c}{S_f}$ (C), the binary entropy S_{bin} (D) and the binary entropy contribution over the full one $\frac{S_{bin}}{S_f}$ (E). The black markers represent common concepts, while the other ones are indicated in orange. The data shown here correspond to the concepts in the physics corpus of section 2.1

features are constrained, as explained in subsection 1.2.1. The maximum entropy distribution is then the maximally unbiased one that fulfills these features derived from the data. In such a way, the prescribed constraints dictate the functional form of the maximum entropy distribution. Operationally, in order to enforce the constraints derived from the expected features, we adopt the Lagrangian multipliers formalism (see subsection 1.2.1, Equation 1.48 and Equation 1.51). In the rest of the Section, we characterize two maximum entropy models, detailing the calculations that lead to the associated distributions from the respective constraints.

A.1.2.1 Discrete tf

First, we consider the case where the average term-frequency of a concept c, $\langle tf_c \rangle$, is imposed as a constraint. The probability mass function describing the expected frequency of a concept c that appears k times is denoted as $q_c(k)$. The system of equations to solve in order to maximize the

entropy is

$$S_{\langle tf_c\rangle} = -\sum_{k=1}^{\infty} q_c(k)\ln q_c(k) - \lambda\left(\sum_{k=1}^{\infty} k\, q_c(k) - \langle tf_c\rangle\right) - \nu\left(\sum_{k=1}^{\infty} q_c(k) - 1\right). \tag{A.2}$$

The Lagrange multipliers ν and λ are associated to the normalization condition and the constraint on $\langle tf_c\rangle$, respectively. The probability mass function is then The maximization of Equation A.2 with respect to $q_c(k)$ is performed as $\frac{\partial S_{\langle tf_c\rangle}}{\partial q_c(k)} = 0$, which gives:

$$-\ln q_c(k) - 1 - \lambda k - \nu = 0\,. \tag{A.3}$$

Thus, the probability mass function q_c reads

$$q_c(k) = e^{-(\nu+1)-\lambda k}\,. \tag{A.4}$$

The probability mass function is an exponential with parameter λ where the constant term $e^{-(\nu+1)}$ is associated to the normalization condition. To impose such constraint, we maximize Equation A.2 with respect to ν, *i.e.* the Lagrange multiplier associated to that constraint, as $\frac{\partial S_{\langle tf_c\rangle}}{\partial \nu} = 0$ obtaining

$$\begin{aligned}\sum_{k=1}^{\infty} q_c(k) = e^{-(\nu+1)}\sum_{k=1}^{\infty} e^{-\lambda k} = e^{-(\nu+1)}\frac{e^{-\lambda}}{1-e^{-\lambda}} = 1\,,\\ e^{-(\nu+1)} = \frac{1-e^{-\lambda}}{e^{-\lambda}}\,,\end{aligned} \tag{A.5}$$

where the last identity in the first line is obtained from the geometric series

$$\sum_{k=1}^{\infty} r^k = \frac{r}{1-r}\,. \tag{A.6}$$

Likewise, the constraint on $\langle tf_c\rangle$ is enforced by the maximization of Equation A.2 with respect to λ, $\frac{\partial S_{\langle tf_c\rangle}}{\partial \lambda} = 0$, as

$$\begin{aligned}\sum_{k=1}^{\infty} k\, q_c(k) = e^{-(\nu+1)}\sum_{k=1}^{\infty} k\, e^{-\lambda k} = -e^{-(\nu+1)}\frac{\partial}{\partial\lambda}\sum_{k=1}^{\infty} e^{-\lambda k} = \frac{1}{1-e^{-\lambda}} = \langle tf_c\rangle\,,\\ e^{-\lambda} = 1 - \frac{1}{\langle tf_c\rangle}\,.\end{aligned} \tag{A.7}$$

Substituting the probability mass function of Equation A.4 in the definition of the entropy gives, for every concept,

$$
\begin{aligned}
S_{\langle tf_c \rangle} &= -\sum_{k=1}^{\infty} q_c(k) \ln q_c(k) = 1 + \nu + \lambda \langle tf_c \rangle = -\ln\left(\frac{1}{\langle tf_c \rangle \left(1 - \frac{1}{\langle tf_c \rangle}\right)}\right) - \langle tf_c \rangle \ln\left(1 - \frac{1}{\langle tf_c \rangle}\right) \\
&= -\ln\left(\frac{1}{\langle tf_c \rangle \left(1 - \frac{1}{\langle tf_c \rangle}\right)}\right) - \langle tf_c \rangle \ln\left(1 - \frac{1}{\langle tf_c \rangle}\right) \\
&= -\ln\left(\frac{1}{\langle tf_c \rangle}\right) + \ln\left(1 - \frac{1}{\langle tf_c \rangle}\right) - \langle tf_c \rangle \ln\left(1 - \frac{1}{\langle tf_c \rangle}\right) \\
&= -\ln\left(\frac{1}{\langle tf_c \rangle}\right) - \left(\langle tf_c \rangle - 1\right) \ln\left(1 - \frac{1}{\langle tf_c \rangle}\right) \\
&= -\langle tf_c \rangle \left(\frac{1}{\langle tf_c \rangle} \ln\left(\frac{1}{\langle tf_c \rangle}\right) + \left(1 - \frac{1}{\langle tf_c \rangle}\right) \ln\left(1 - \frac{1}{\langle tf_c \rangle}\right)\right).
\end{aligned}
$$

Finally, simple algebra from the last line gives the formula reported in Equation 2.6.

A.1.2.2 Density of tf

The second maximum entropy model is designed to describe the rescaled version of the term-frequency distribution of a concept c. Such term-frequency density is defined as $rtf_c(\alpha) = \frac{tf_c(\alpha)}{L(\alpha)}$, where $L(\alpha)$ is the length of document α calculated as the number of words. The term-frequency density, indeed, is better suited to describe the relevance of a concept within documents in the case of a length distribution which is inhomogeneous. In the opposite case, *i.e.* when documents have the same length, the (discrete) term-frequency distribution and the rescaled version are identical up to a scaling factor. The analytical expression of the probability density function $p_c(x)$ is determined by maximizing its entropy S_{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:

$$
\begin{aligned}
\tilde{S} = &-\int_0^{\infty} p_c(x) \ln p_c(x)\, dx \\
&+ \gamma \left(\langle \ln(rtf_c) \rangle - \int_0^{\infty} \ln(x)\, p_c(x)\, dx \right) \\
&+ \eta \left[\sigma^2(\ln(rtf_c)) - \int_0^{\infty} \left(\ln(x) - \int_0^{\infty} \ln(x)\, p_c(x)\, dx \right)^2 p_c(x)\, dx \right] \\
&+ \nu \left(1 - \int_0^{\infty} p_c(x)\, dx \right),
\end{aligned}
\tag{A.8}
$$

where γ, η and ν are the Lagrange multipliers associated to the constraints $\langle \ln(rtf_c) \rangle$, $\sigma^2(\ln(rtf_c))$ and the normalization condition of $p_c(x)$, respectively. Maximizing Equation A.8 with respect to

$p_c(x)$, $\frac{\partial \tilde{S}}{\partial p_c(x)} = 0$, we obtain:

$$-\ln p_c(x) - 1 - \gamma \ln(x) - \eta \left(\ln(x) - \mu\right)^2 - \nu = 0\,, \tag{A.9}$$

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

$$p_c(x) = \frac{\mathrm{e}^{-(\nu+1)} \mathrm{e}^{-\eta\left(\ln(x)-\mu\right)^2}}{x^\gamma}\,. \tag{A.10}$$

As we have done in the previous case, subsubsection A.1.2.1, we must impose the normalization condition on the probability density function Equation A.10. Moreover, we also compute the parameters η and γ, similarly to what we performed in Equation 2.13 and Equation 2.14. Since we have already detailed the process to calculate the parameters, we report here directly the final expression of the probability density function:

$$p_c(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma\, x} \exp\left[-\frac{(\ln x - \mu)^2}{2\,\sigma^2}\right] \qquad \text{with } x > 0\,. \tag{A.11}$$

Such probability density function corresponds to the lognormal reported in Equation 2.18, where the parameters μ and σ^2 are computed directly from the empirical constraints on the average and the variance of the logarithm of x:

$$\mu = \int_0^\infty \ln(x)\, p_c(x)\, dx \equiv \langle \ln(r\,t\,f_c) \rangle\,, \quad \sigma^2 = \int_0^\infty \left(\ln(x) - \mu\right)^2 p_c(x)\, dx \equiv \sigma^2(\ln(r\,t\,f_c))\,. \tag{A.12}$$

Note that μ is a constant fully determined from the data and is not a function of σ^2. The maximum entropy S_{max} associated to the probability density function in Equation A.11 is then:

$$S_{max} = -\int_0^\infty p_c(x) \ln p_c(x)\, dx = \ln\left(\sqrt{2\pi}\,\sigma\right) + \mu + \frac{1}{2}\,. \tag{A.13}$$

A.1.3 Equivalence between the Kullback-Leibler divergence and the residual entropy

In section 2.2, we introduced the definition of *residual entropy* of a concept c, $S_d(c)$, as the difference between its maximum entropy, $S_{max}(c)$, and the conditional one, $S_c(c)$. Here we show that the residual entropy, S_d, which measures the "generality" of concepts, is exactly equivalent to the Kullback-Leibler divergence (KL), a widely adopted to rigorously compute the deviation between two probability distributions [185]. More specifically, we consider the KL between the maximum entropy probability distribution, q_c, and the empirical one, p_c. For the sake of simplicity, we prove the equivalence with the residual entropy in the case where p and q are probability mass functions representing discrete random variables, although the same reasoning can be used for probability density functions. In particular, we remind that $p_c(k)$ describes the

observed probability that the term-frequency of a concept c, tf_c, is equal to k, thus $p_c(k) = \frac{N(k)}{N_a}$. The KL from q to p is defined as:

$$D_{\mathrm{KL}}(p||q) = \sum_{k=1}^{\infty} p_c(k) \ln\left(\frac{p_c(k)}{q_c(k)}\right) = -\sum_{k=1}^{\infty} p_c(k) \ln q_c(k) + \sum_{k=1}^{\infty} p_c(k) \ln p_c(k)\,. \tag{A.14}$$

The last term in the (A.14) is nothing else, apart for the sign, than the conditional entropy S_c:

$$S_c = -\sum_{k=1}^{\infty} p_c(k) \ln p_c(k)\,. \tag{A.15}$$

The first term, instead, can be rewritten using the maximum entropy probability q_c (see Equation 2.12) as:

$$\begin{aligned}
-\sum_{k=1}^{\infty} p_c(k) \ln q_c(k) &= -\sum_{k=1}^{\infty} p_c(k) \ln\left(\frac{\frac{\mathrm{e}^{-\lambda k}}{k^s}}{\mathrm{Li}_s(\mathrm{e}^{-\lambda})}\right) \\
&= \sum_{k=1}^{\infty} p_c(k) \ln\left[\mathrm{Li}_s(\mathrm{e}^{-\lambda})\right] - \sum_{k=1}^{\infty} p_c(k) \ln\left(\frac{\mathrm{e}^{-\lambda k}}{k^s}\right) \\
&= \ln\left[\mathrm{Li}_s(\mathrm{e}^{-\lambda})\right] - \sum_{k=1}^{\infty} p_c(k) \ln\left(\frac{\mathrm{e}^{-\lambda k}}{k^s}\right) \\
&= \ln\left[\mathrm{Li}_s(\mathrm{e}^{-\lambda})\right] + \lambda \sum_{k=1}^{\infty} p_c(k)\, k + s \sum_{k=1}^{\infty} p_c(k) \ln k \\
&= \ln\left[\mathrm{Li}_s(\mathrm{e}^{-\lambda})\right] + \lambda \langle k \rangle + s \langle \ln k \rangle \equiv S_{max}\,.
\end{aligned} \tag{A.16}$$

Plugging the results of Eqs. (A.15) and (A.16) into (A.14) we get:

$$D_{\mathrm{KL}}(q||p) = -\sum_{k=1}^{\infty} q(k) \ln p(k) + \sum_{k=1}^{\infty} q(k) \ln q(k) = S_{max} - S_c\,. \tag{A.17}$$

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

A.1.4 Comparisons between sets

A.1.4.1 Comparison between concept rankings based upon S_d and IDF

Despite both the nverse document frequency, IDF, and the residual entropy, S_d, are defined on the corpus of documents under scrutiny, they actually encode different information. The first penalizes concepts that are frequently present across articles, while the latter is intrinsically related to the term-frequency distribution of a concept. Therefore, we may expect to observe some correlation between them although they are different quantities. To compare the list of concepts ranked upon each quantity, we calculate the overlap, $O_{n,m}$, between the set of concepts

in the nth percentile slice[1] of the IDF, $\mathcal{A}_n$, and the set of concepts in the mth percentile slice of S_d, $\mathcal{B}_m$. Thus, we have:

$$O_{n,m} = \frac{|\mathcal{A}_n \cap \mathcal{B}_m|}{|\mathcal{A}_n|}, \tag{A.18}$$

with $O_{n,m} \in [0,1]$. As usual, $O_{n,m} = 1$ denotes that the two sets share exactly the same elements, *i.e.* completely overlap, while $O_{n,m} = 0$ characterizes two sets with no elements in common.

A.2 Datasets

In this section, we provide additional findings from the analysis of the physics corpus in subsection A.2.1.Moreover, we also describe supplementary findings related to the corpus of web documents about climate change in subsection A.2.2. We recall that we used the maximum entropy model based on the (discrete) term-frequency tf_c for the former corpus, while for the latter we adopted term-frequency density, rtf_c.

A.2.1 Physics

A.2.1.1 Similarity between partitions

In order to quantify the similarity between the partitions obtained in different runs, we compute the Normalized Mutual Information (NMI) and the Normalized Variation of Information (NVI) between the partitions for any pair of runs, as defined in subsection 1.1.5. In Figure A.2, we display for every filtering percentile from $p = 10$ to $p = 50$ the histograms of the NMI and the NVI. Moreover, we show the histogram of the modularity Q, defined in subsection 1.1.4, for all the runs. We observe that the modularity tend to increase significantly with the filtering percentile p. On the contrary, both the NMI and NVI does not show a remarkable shift towards higher or lower values as p increases, but they tend to be narrower distributed close to the central peak.

A.2.1.2 Jaccard score between communities and `arXiv` categories

The Jaccard scores displayed in Figure A.3 as an heatmap aims to detail the correspondence between the communities of articles detected in the similarity networks and the categories to whom these articles belong. Indeed, we observe the presence of strong analogies between the communities of articles and the `arXiv` categories.

[1] Given a probability distribution $p(x)$, the kth *slice* of the percentile contain those values x which are included in the kth percentile, $\tilde{P}_k$, but are not present in the $(k-10)$th percentile, *i.e.* $x \in \left[\tilde{P}_{k-10}, \tilde{P}_k\right]$, $k \in \{10, 20, \ldots, 90\}$.

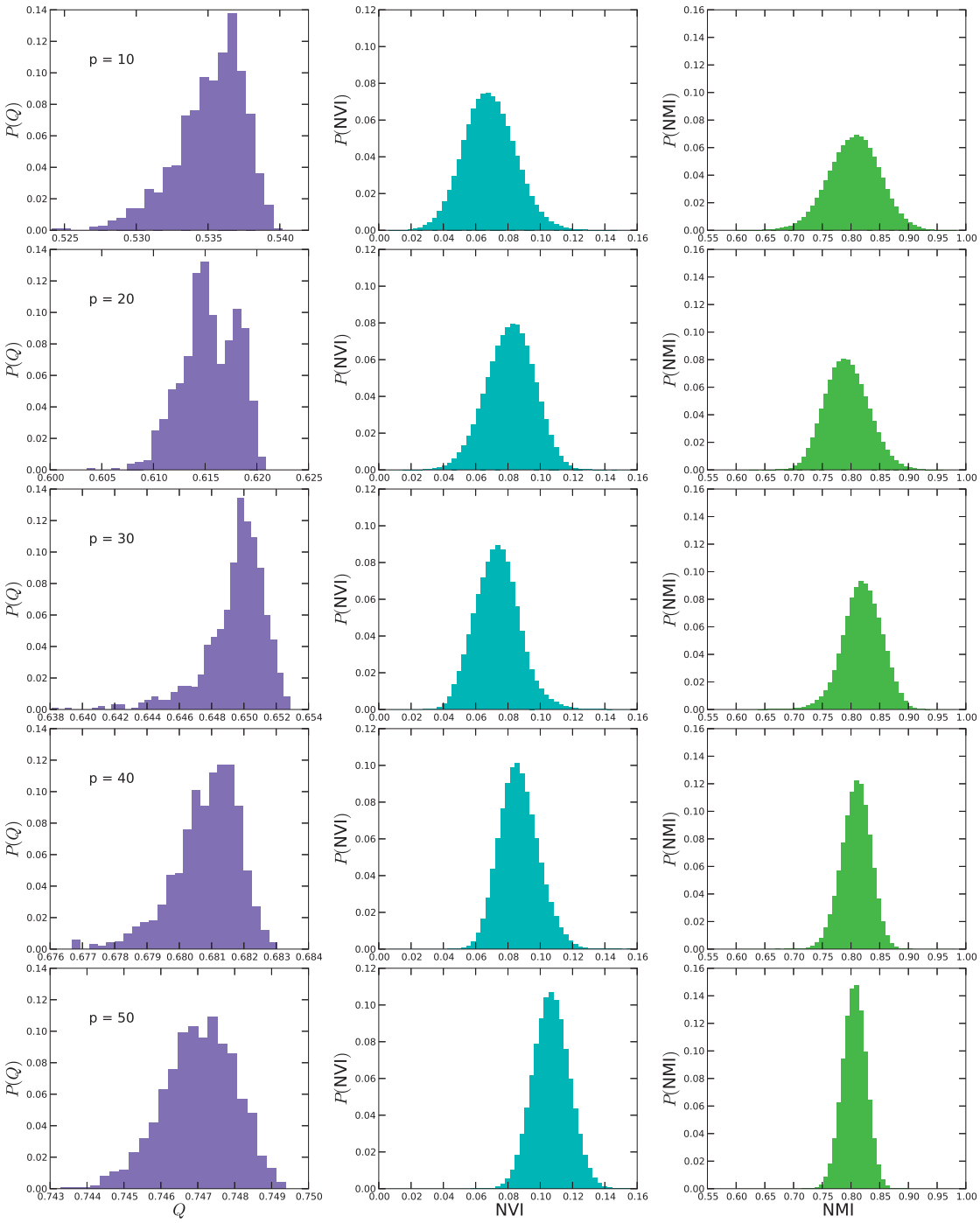

Figure A.2 – Histograms of the modularity Q (left column) calculated over all the 1000 runs of the Louvain method at each filtering percentile p. In addition, the histograms of the NMI (center) and NVI (right) between any pair of partitions are shown.

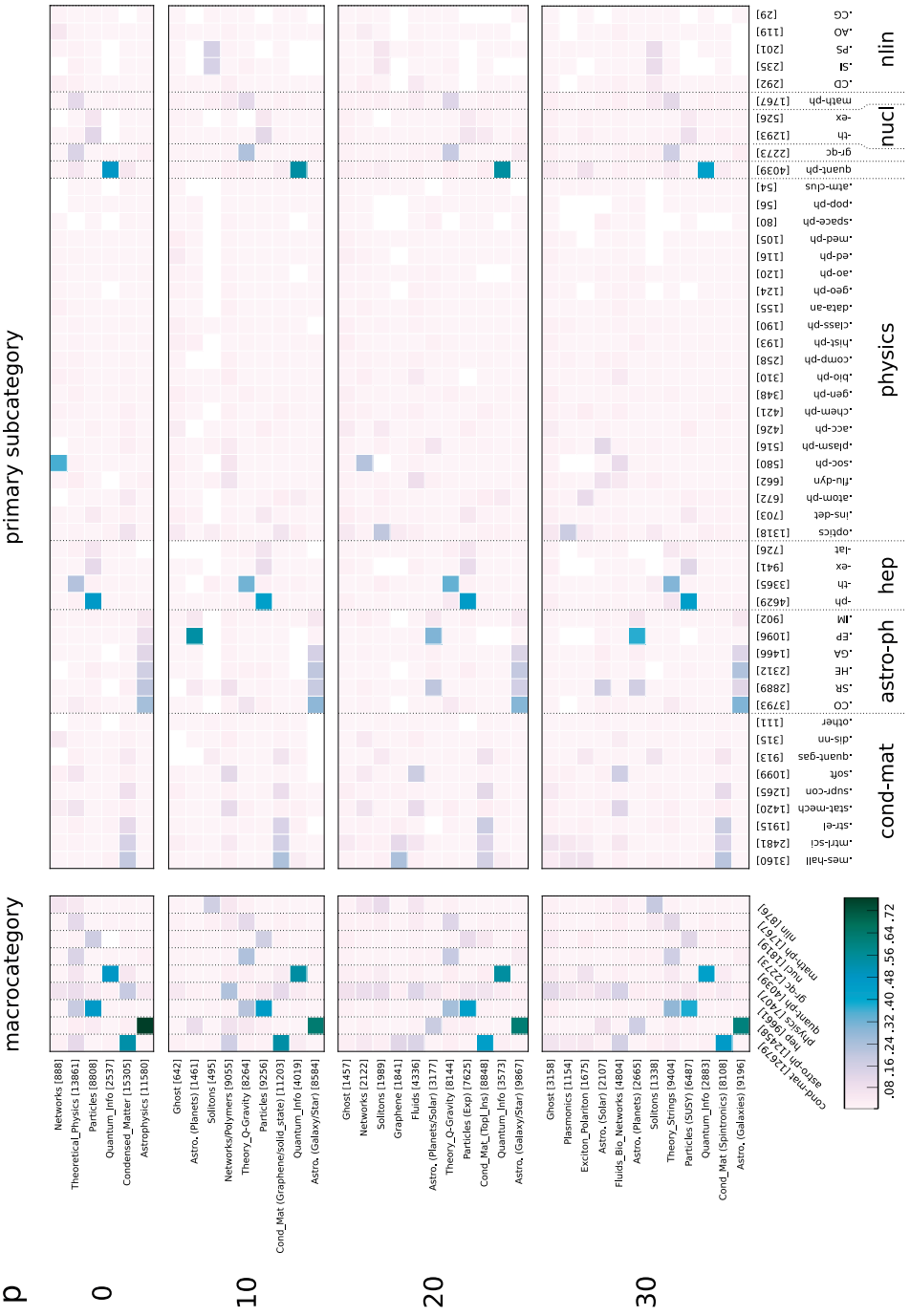

Figure A.3 – Jaccard score between the communities of articles and the categories of `arXiv`. Each row denotes a filtering percentile, p, where the communities are identified with the same label of the Sankey diagram in Figure 2.12. The number in square brackets denotes the size of the community or category. The columns of the heatmaps on the left correspond to the macrocategories, as defined in Table 2.1, sorted by decreasing size. The columns of the heatmaps on the right represent the subcategories, sorted within the same macrocategory by decreasing size. The vertical dashed lines delimit the macrocategories.

A.2.1.3 Differences between S_d and IDF rankings

In order to examine the correlation between the ranking of concepts based upon the residual entropy, S_d, and the ranking based upon the IDF, we calculate the overlap score O of the sets of concepts belonging to the percentile slices of these quantities, as defined in Equation A.18. The result is shown in Figure A.4: in the case of the 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 in ascending order of S_d (thus from the most generic to the least one). According to the definition of O, matrices are normalized by row. The analysis of the overlap matrix denotes 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 across the collection, albeit this is not always the case.

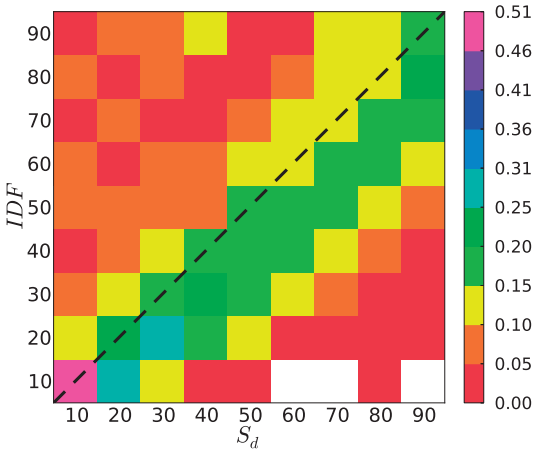

Figure A.4 – Overlap between sets of concepts ranked according to the residual entropy S_d and IDF for the physics corpus. The color of the cells denotes the value of overlap O where the white corresponds to the absence of overlap. Matrices are normalized by row and the dashed line indicates the main diagonal.

A.2.1.4 Comparison of the communities obtained after filtering concepts on S_d and IDF

Given the results of the overlap between the sets of concepts ranked according to S_d and IDF, we compare the community structure uncovered after filtering concepts with each of two criteria. In Figure A.5 we report the heatmaps of the Jaccard score computed between the communities of the similarity networks obtained by pruning out a given fraction p 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, we can see that there is always a certain degree of similarity between 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 that using residual entropy to filter concepts is not equivalent to the filter based on the inverse document frequency, which is the most widespread method to remove concepts in order to sparsify the network of similarity between articles.

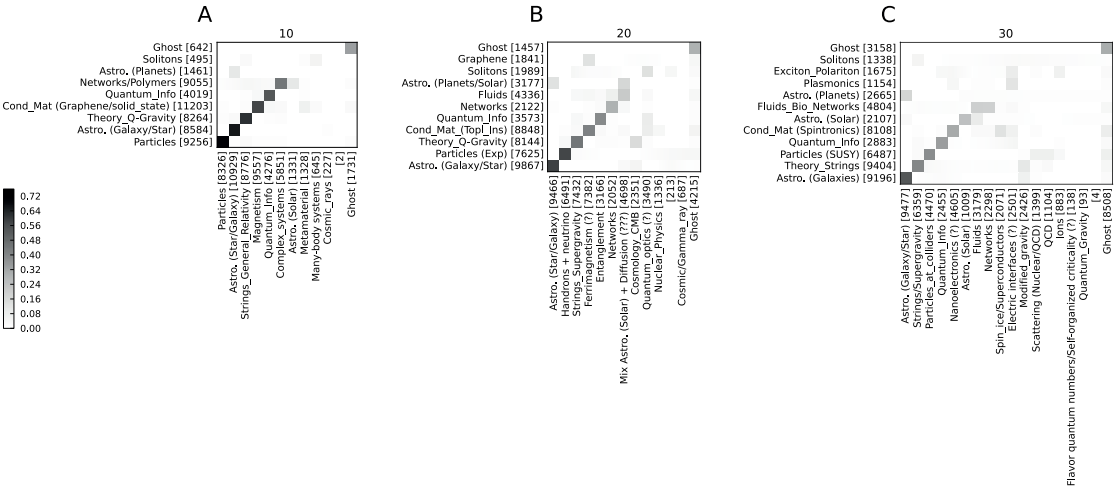

Figure A.5 – Jaccard scores among the communities of the similarity networks obtained after filtering concepts either using entropy (y-axis) or IDF (x-axis). Each label accounts for the main topic and the size of a give community. Each matrix refers to removing, respectively, the 10% (A), 20% (B) and 30% (C) of the concepts. The values of the score 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 score 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 score 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 score which is 0.30.

A.2.1.5 Ranking of concepts within papers

In the following, we want to understand if the entropic selection of concepts could be adopted also to rank concepts and, in turns, if we can use those rankings to classify papers. To this aim, we select ten highly cited papers from the physics corpus, which are reported in Table A.1. For each of them we consider the 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 top ten concepts of each ranking are listed in Table A.2. Qualitatively speaking, the ranked

lists of concepts presented in 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 this task. However, these conclusions are not surprising: both S_d and IDF are global measures, defined on the whole corpus in order to quantify the importance of a concept for the entire corpus. Conversely, both TF and TF-IDF are local measures that capture the significance of a concept within papers.

arXiv ID	# cit	Prim. cat.	Secondary category(ies)	Venue
1306.5856	1572	cond-mat.mtrl-sci	–	*Nat Nanotech* **8**, 235–246 (2013)
1301.0842	390	astro-ph.EP	–	*Astroph Jour* **766**, 81 (2013)
1308.0321	478	cond-mat.quant-gas	cond-mat.str-el, quant-ph	*Phys Rev Lett* **111**, 185301 (2013)
1301.1340	272	hep-ph	–	*Rep Prog Phys* **76**, 056201 (2013)
1301.0319	415	astro-ph.SR	astro-ph.IM	*Astrophys J Suppl Ser* **208**, 4 (2013)
1304.6875	667	astro-ph.HE	astro-ph.SR, cond-mat.quant-gas, gr-qc	*Science* **340**, Issue 6131 (2013)
1306.2314	254	astro-ph.CO	–	*Phys Rev D* **88**, 043502 (2013)
1311.6806	231	astro-ph.EP	–	*PNAS* **110**, 19175 (2013)
1302.5433	195	cond-mat.supr-con	cond-mat.mes-hall	*J Phys Conden. Matter* **25**, 233201 (2013)
1303.3572	26	cond-mat.str-el	hep-ph, quanth-ph	*Phys Rev B* **89**, 045127 (2014)

Table A.1 – 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 primary category and eventual secondary ones. Finally, we provide the publication venue. The number of citations has been retrieved from [284] the 19th of December 2017.

Table A.2 – List of the ten most generic concepts for the papers listed in Table A.1. We rank concepts using: residual entropy S_d, inverse document frequency IDF, term-frequency TF and TF-IDF. Concepts indicated as common by SW are marked with 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 *	Phonon	Phonon
	Regularization	Measurement *	Graphene	Graphene
	Intensity	Field *	Electron *	Graphite
	Temperature *	Potential *	Energy *	Raman spectroscopy
	Field *	Mass *	Graphite	Raman scattering
	Optics *	Particles *	Resonance *	Electron *
	Electromagnet *	Temperature *	Frequency *	Carbonate *
	Energy *	Probability *	Measurement *	Resonance *
	Mass *	Units *	Scattering *	Wave vector
	Wavelength *	Vector *	Intensity	Selection rule
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

Continued on next page

continued from previous page				
arXiv ID	S_d	IDF	TF	TF-IDF
	Space telescopes	Potential *	Kepler Objects of Interest	False positive rate
	Temperature *	Mass *	Periodate *	Star
	Statistical error	Temperature *	Frequency *	Eclipsing binary
	Field *	Probability *	Signal to noise ratio	Eclipses
	Optics *	Units *	False positive rate	Signal to noise ratio
	Mass *	Frequency *	Eclipsing binary	Neptune
	Frequency *	Periodate *	Eclipses	Earth-like planet
	Fluctuation *	Velocity *	Orbit *	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
	Factorisation	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 of the exression *	Field *	Frequency *	Massive stars
	Regularization	Potential *	White dwarf	Stellar evolution
	Intensity	Mass *	Angular momentum *	Angular momentum *
Continued on next page				

continued from previous page				
arXiv ID	S_d	IDF	TF	TF-IDF
	Temperature *	Particles *	Massive stars	Mass *
	Field *	Temperature *	Temperature *	Planet
	Optics *	Probability *	Pressure *	Red supergiant
	Energy *	Units *	Luminosity	Asteroseismology
	Mass *	Electron *	Stellar evolution	Zero-age main 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
	Mass *	Vector *	Companion	Binary star
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 *	Kepler Objects of Interest	Signal to noise ratio
	Intensity	Potential *	Light curve	Light curve
	Temperature *	Mass *	Photometry	Habitable zone
	Statistical error	Temperature *	Eclipses	Photometry
	Energy *	Probability *	Stellar radii	Stellar radii
	Mass *	Periodate *	Extrasolar planet	Eclipses
	Wavelength *	Universe *	Habitable zone	Eclipsing binary
Continued on next page				

continued from previous page				
arXiv ID	S_d	IDF	TF	TF-IDF
	Fluctuation *	Objective *	Eclipsing binary	Extrasolar planet
	Uniform distribution	Statistics	Event *	High resolution échelle spectrometer
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
	Right Hand Side 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 of the exression *	Energy *	Bosonization	Dyon
	Regularization	Field *	Dyon	Electromagnetism
	Strong interactions	Potential *	Charge *	U(1) *
	Degree of freedom	Mass *	Condensation	Witten effect
	Vacuum *	Particles *	Fermion *	Bosonization
	Field *	Units *	Statistics	Condensation
	Electromagnet *	Vector *	U(1) *	Projective construction
	Energy *	Periodate *	Electromagnetism	Time-reversal symmetry
	Mass *	Symmetry *	Symmetry *	Fermion *
	Fluctuation *	Statistics	Time-reversal symmetry	Mean field

Table A.2 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 Table A.3 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 subsection 2.3.1. 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. However, this phenomenon has already been observed. More importantly, we clearly see how the entropic filtering removes also concepts classifiable as generic that have not been marked as such by SW.

Table A.3 – List of the ten most generic concepts per paper in Table A.1 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 SW 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			
	Experimental data *	Electronic 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 *	Diffraction *	Backscattering	Superlattice
	Optics *	Nanostructure	Scanning tunneling microscope	Van Hove singularity
	Electromagnet *	Hydrostatics	Graphite	Surface plasmon
	Energy *	Electron scattering	Nitriding	Exciton
	Mass *	Circular polarization *	Dirac point	Nanomaterials
	Wavelength *	Space-time singularity	Normal mode	Intervalley 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	Asteroseismology
	Statistical error	Spectrographs	Star counts	High accuracy radial velocity planetary search
	Field *	Angular distance	Earth-like planet	Hot Jupiter
	Optics *	Stellar magnitude	Orbital elements	Triple system
	Mass *	Extinction	Eccentricity	Neptune
	Frequency *	Kolmogorov-Smirnov test	Primary stars	Periastron
	Fluctuation *	Solar neighborhood	Giant planet	Planet

Continued on next page

Appendix A. Entropic selection of concepts in networks of similarity between articles

continued from previous page				
arXiv ID	$p = 0$	$p = 10$	$p = 20$	$p_{opt} = 30$
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 interaction	Clebsch-Gordan coefficients
	Field *	Baryon number	Renormalisation group equations	Neutrino telescope
	Electromagnet *	Subgroup	CP violation	CP violating phase
	Energy *	Permutation	Euler angles	Proton decay
	Mass *	Quark mass	Rotation group *	Neutrino mixing angle
	Equation of motion *	Irreducible representation	Superpotential	Complex conjugate representation
	Momentum *	Embedding	Reactor neutrino experiments	Neutralino
1301.0319	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 of the exression *	Low-mass stars	Circumstellar envelope	Radiative Diffusion
	Regularization	Massive stars	Early-type star	Optical bursts

Continued on next page

continued from previous page

arXiv ID	$p = 0$	$p = 10$	$p = 20$	$p_{opt} = 30$
	Intensity	Diffusion coefficient	Helium shell flashes	Asteroseismology
	Temperature *	Accretion disk	Modified gravity	Stellar oscillations
	Field *	Irradiance	Evolved stars	Zero-age main sequence stars
	Optics *	Sloan Digital Sky Survey	Neutron star	Classical nova
	Energy *	Viscosity	Hertzsprung-Russell diagram	Large Synoptic Survey Telescope
	Mass *	Hydrostatics	Supernova	Giant branches
1304.6875	A Massive Pulsar in a Compact Relativistic Binary			
	Order of magnitude *	Accretion	Radio telescope	Lunar Laser Ranging experiment
	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 Gravitational-Wave Observatory
	Optics *	Spectrographs	Roche Lobe	Radiation damping
	Energy *	Space-time singularity	Mass accretion rate	Barycenter
	Mass *	Stellar surfaces	Peculiar velocity	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	Cutoff scale	Mass function	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	Tellurate	Halo finding algorithms
	Optics *	Wavenumber *	Monte Carlo Markov chain	Baryon acoustic oscillations

Continued on next page

continued from previous page

arXiv ID	$p = 0$	$p = 10$	$p = 20$	$p_{opt} = 30$
	Mass *	Absorption feature	Cold dark matter	Void
	Fluctuation *	Flavour	Stellar feedback	Strong gravitational lensing
1311.6806	Prevalence of Earth-size planets orbiting Sun-like stars			
	Stefan-Boltzmann constant	Simulations *	Hertzsprung-Russell diagram	Eclipsing binary
	Solar mass	Companion	Earth-like planet	Asteroseismology
	Intensity	Stellar surfaces	Monte Carlo Markov chain	Limb darkening
	Temperature *	Stellar magnitude	Hydrogen 21 cm line	Planet
	Statistical error	Host star	Keck Array	High resolution échelle spectrometer
	Energy *	Orbit Eccentricity	Parallax	Eclipses
	Mass *	Droplet *	Eclipsing binary	Habitable 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 of the exression *	Rest mass *	Topological insulator	Chern number
	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

Continued on next page

continued from previous page				
arXiv ID	$p = 0$	$p = 10$	$p = 20$	$p_{opt} = 30$
1303.3572	3-dimensional bosonic topological insulators and its exotic electromagnetic response			
	Right Hand Side of the exression *	Dirac fermion	Band insulator	Hall conductance
	Regularization	Quantum Hall Effect	Topological insulator	Electric magnetic
	Strong interactions	SU(2) *	Mott insulator	Long-range entanglement
	Degree of freedom	Effective field theory	Charge conservation	Topological field theory
	Vacuum *	Parton	Magnetic monopole	Axion
	Field *	Deconfinement	Edge excitations	Exciton
	Electromagnet *	Screening effect	Topological order	Short-range entanglement
	Energy *	Effective Lagrangian	Berry phase	Symmetry protected topological order
	Mass *	Directional derivative	Fractional charge	Group cohomology
	Fluctuation *	Global symmetry	Electromagnetism	Charge quantization

The information presented in Table A.3 confirms the power of our filtering methodology. In analogy to what we have done in Table A.2, we check if S_d still outperforms other rankings also in the filtered networks. For this reason, in Table A.4 we report the rankings of the concepts at the optimal level of filtering ($p_{opt} = 30\%$). A quick glance at its columns indicates that, albeit being more specific, concepts ranked using S_d are still capable of describing the content of the document, and such ranking is more dissimilar to the others three.

Table A.4 – Ten most generic concepts among those available at the optimal level of filtering, $p_{opt} = 30\%$, for S_d. Columns are the same as Table A.2, reporting the ranking upon different quantities. The highlighted columns are those corresponding to S_d and TF-IDF which represent the best rankings.

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			
	Monochromator	Exciton	Surface enhanced Raman spectroscopy	Surface enhanced Raman spectroscopy
	Surface plasmon resonance	Superlattice	Van Hove singularity	Kohn anomaly
	Bilayer graphene	Graphene layer	Grüneisen parameter	Grüneisen parameter
Continued on next page				

continued from previous page

arXiv ID	$S_d(p_{opt}=30)$	IDF ($p_{opt}=30$)	TF ($p_{opt}=30$)	TF-IDF ($p_{opt}=30$)
	Graphene layer	Bilayer graphene	Kohn anomaly	Van Hove singularity
	Superlattice	Surface plasmon	Hexagonal boron nitride	Hexagonal boron nitride
	Van Hove singularity	Monochromator	Nanocrystalline	Nanocrystalline
	Surface plasmon	Van Hove singularity	Graphene layer	Graphene layer
	Exciton	Nanocrystal	Exciton	Fullerene
	Nanomaterials	S-process	Nanocrystal	Nanocrystal
	Intervalley 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
	Neptune	Neptune	Triple system	Triple system
	Periastron	Asteroseismology	White dwarf	Logarithmic distribution
	Planet	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
	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

Continued on next page

continued from previous page

arXiv ID	$S_d(p_{opt}=30)$	IDF ($p_{opt}=30$)	TF ($p_{opt}=30$)	TF-IDF ($p_{opt}=30$)
	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 for Neutrino Oscillation	Solar neutrino
	Clebsch-Gordan coefficients	Mantle	Tri Bimaximal mixing	Reactor Experiment for Neutrino Oscillation
	Neutrino telescope	Two Higgs Doublet Model	Clebsch-Gordan coefficients	Trimaximal mixing
	CP violating phase	Supermultiplet	Super-Kamiokande	SNO+
	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	Mikheev-Smirnov-Wolfenstein effect
	Neutralino	Massive neutrino	Type I seesaw	Cabibbo Angle
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	Asteroseismology
	Asteroseismology	Pre-main-sequence star	Asteroseismology	Zero-age main sequence stars
	Stellar oscillations	Asymptotic giant branch	Pre-main-sequence star	Pre-main-sequence star
	Zero-age main sequence stars	Zero-age main sequence stars	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

Continued on next page

continued from previous page

arXiv ID	$S_d(p_{opt}=30)$	IDF ($p_{opt}=30$)	TF ($p_{opt}=30$)	TF-IDF ($p_{opt}=30$)
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 Gravitational-Wave 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 Gravitational-Wave Observatory	Solar system barycenter	Radio pulsar
	Space velocity	Millisecond pulsar	VLT telescope	Dispersion measure
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	WDM particles	Ultraviolet background
	Free streaming	Gamma ray burst	Free streaming	Quasar
	Quasar	Void	Ultraviolet background	Free streaming
	Active Galactic Nuclei	Baryon acoustic oscillations	Redshift bins	Redshift bins
	Planck data	Reionization	Temperature-density relation	Temperature-density relation
	Halo finding algorithms	Satellite galaxy	Reionization	Effective optical depth
	Baryon acoustic oscillations	Nuisance parameter	Warm dark matter	Warm dark matter
	Void	Population III	Effective optical depth	Reionization

Continued on next page

continued from previous page				
arXiv ID	$S_d(p_{opt}=30)$	IDF ($p_{opt}=30$)	TF ($p_{opt}=30$)	TF-IDF ($p_{opt}=30$)
	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	Habitable zone	Eclipses
	Planet	Limb darkening	Eclipsing binary	Eclipsing binary
	High resolution échelle spectrometer	Mars	Planet	High resolution échelle spectrometer
	Eclipses	Asteroseismology	Mars	Mars
	Habitable zone	Habitable zone	High resolution échelle spectrometer	Horizon Run simulation
	Gaussian process	Gaussian process	Limb darkening	Planet
	Mars	Ephemerides	False positive rate	False positive rate
	Orange dwarf	High resolution échelle spectrometer	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	P-wave	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	Moore-Read Pfaffian wavefunction
	Weak antilocalization	Weyl fermion	Non-Abelian statistics	Majorana zero mode
	Non-Abelian statistics	Non-Abelian statistics	Moore-Read Pfaffian wavefunction	Non-Abelian statistics
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	Projective construction
Continued on next page				

continued from previous page				
arXiv ID	$S_d(p_{opt}=30)$	IDF ($p_{opt}=30$)	TF ($p_{opt}=30$)	TF-IDF ($p_{opt}=30$)
	Topological field theory	Topological field theory	Group cohomology	Group cohomology
	Axion	Electric magnetic	Topological field theory	Topological field theory
	Exciton	Long-range entanglement	Exciton	Response theory
	Short-range entanglement	Symmetry protected topological order	Response theory	Charge quantization
	Symmetry protected topological order	Dyon	Charge quantization	Short-range entanglement
	Group cohomology	Short-range entanglement	Axion	Symmetry protected topological order
	Charge quantization	Charge quantization	Hall conductance	Long-range entanglement

A.2.2 Climate change web documents

A.2.2.1 Entropic filtering

The conditional and maximum entropy calculated for the term-frequency density of the keywords are outlined in subsubsection A.1.2.2. The position of points in the $(S_c, S_{\max})$ plane is reported in Figure A.6.

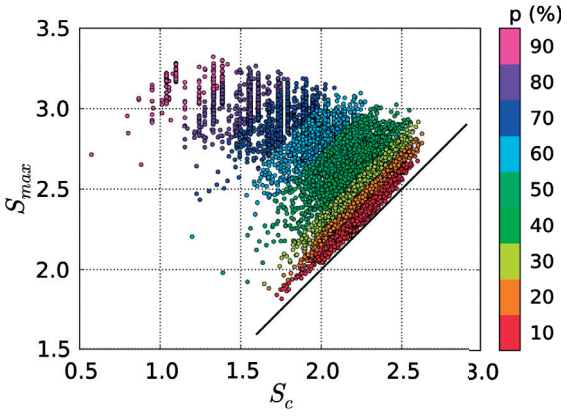

Figure A.6 – Relation between the empirical entropy, S_c, and the maximum one, S_{max} for the climate change corpus. The colors of the points encode the various percentiles of the residual entropy S_d to which concepts belong to.

p (%)	N_{con}	N_a	ρ (%)	$\langle k \rangle$	k_{max}	T	$\langle l \rangle$	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

Table A.5 – 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 ρ, the average and maximum degrees, $\langle k \rangle$ and k_{max}, the transitivity T, the average shortest path length $\langle l \rangle$ and the number of connected components M. The minimum edge weight is equal to $w_{min} = 0.01$.

As in the case of physics concepts, we clearly observe a stratification of the residual entropy S_d on the plane. Examples of the generic concepts found in the percentile slice $p = 10$ are "people", "climate change", "water", "home" and "company". On the other hand, among concepts in 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 Table A.5.

A.2.2.2 Differences between S_d and IDF rankings

Using the same formalism of subsubsection A.2.1.3, the overlap between the sets of concepts ranked alternatively using S_d or IDF is shown in Figure A.7. The heatmap presents a narrow region of high values concentrate on the main diagonal. Compared with the physics corpus, the overlap is much higher, denoting a much stronger relation between the S_d and IDF rankings.

A.3 Numerical implementation with code snippets

In this Section we present a step-by-step description of the algorithms adopted to implement the entropic filtering of concepts. First, we comment the case of the (discrete) term-frequency, where the maximum entropy model is a power-law distribution with a cutoff (subsection A.3.1), and then the case of the term-frequency density, which maximum entropy model is a lognormal

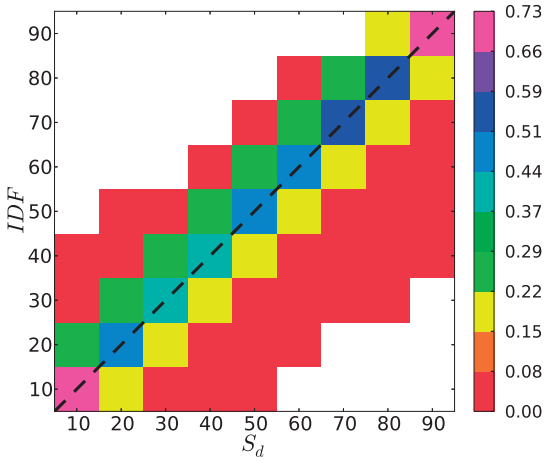

Figure A.7 – Overlap between sets of concepts ranked according to the residual entropy S_d and IDF for the climate change corpus. The matrix is normalized by row and white entries correspond to absence of overlap. The dashed line indicates the main diagonal.

(subsection A.3.2). Our code has been written using the Python programming language [285] making use of several functions available in the `scipy` ecosystem [286].

The core of the filtering method is the comparison between two entropies: the actual/experimental one, S_c, and the expected/theoretical one, $S_{\max}$, associated the maximum entropy principle. Given a corpus of documents $\mathcal{D}$, for each concept c appearing inside a document $\alpha \in \mathcal{D}$, the ScienceWISE platform provides its *boosted term-frequency* $tf_c(\alpha)$, calculated as in Equation 2.1.

A.3.1 Discrete tf

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

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

where $N(x)$ is the number of times the variable X has value x, while M is the total number of values of X. In our case, X is the tf sequence of a concept c and $P(X = x)$ 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 $\langle X \rangle$, σ_X and $\langle \ln(X) \rangle$, respectively: the average, standard deviation and average of the logarithm of 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 construct the experimental term-frequency distribution of the occurrences of concept c.

2. **Extraction of fit parameters:**

The analytical form of the expected power-law with a cutoff is:

$$p(tf_c = k) \equiv p_c^{th}(k) = \frac{1}{Z}\frac{e^{-\lambda k}}{k^s}. \tag{A.19}$$

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

$$Z \equiv \mathrm{Li}_s(\mathrm{e}^{-\lambda}) = \sum_{k=1}^{\infty}\frac{\mathrm{e}^{-\lambda k}}{k^s}, \tag{A.20}$$

The theoretical distribution $p_c^{th}(k)$ depends on two parameters: s and λ. 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 s and λ by solving numerically the following system:

$$\begin{cases} \langle l_{tf}\rangle = & \dfrac{\mathrm{Li}_{s-1}(\mathrm{e}^{-\lambda})}{\mathrm{Li}_s(\mathrm{e}^{-\lambda})}, \\ \\ \langle \ln(l_{tf})\rangle = & \dfrac{-\partial_s \mathrm{Li}_s(\mathrm{e}^{-\lambda})}{\mathrm{Li}_s(\mathrm{e}^{-\lambda})} = \dfrac{\sum_{k=1}^{\infty}\frac{\mathrm{e}^{-\lambda k}}{k^s}\ln(k)}{\mathrm{Li}_s(\mathrm{e}^{-\lambda})}. \end{cases} \tag{A.21}$$

Since the polylogarithm function appears in the above system, we need to use the Python package named `mpmath` [287], which implements functions and methods with arbitrary precision float arithmetics. Thus, we define the two equations that have to the be solved simultaneously as:

```
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 `diff` is used to calculate numerically the derivative of the function `polylog` with respect to s. Then, we use

the `findroot` function of `mpmath` to numerically solve the system of equations with:

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

Typically, the initial values of the parameters are `ci = (0.5,mexp(-0.1))`. The solution of Equation A.21 is stored in `sol`, having s 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 `max_ent` function defined below to compute the maximum entropy.

(b) Use the *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 [197, 288] we use the Python `powerlaw` package to compute the parameters.

3. **Computation of Entropies:**

Given the parameters s and λ, we can compute the maximum entropy of a concept c as:

$$S_{max} = \ln\left[\mathrm{Li}_s(e^{-\lambda})\right] + \lambda \langle tf_c^{exp} \rangle + s \langle \ln\left(tf_c^{exp}\right) \rangle . \tag{A.22}$$

which, implemented in Python, reads as follows:

```
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 computed using Shannon formula (Equation 1.35) from distribution $p_c^{exp}(k)$.

A.3.2 Density of tf

The maximum entropy distribution associated to the case of a rescaled term-frequency sequence, rtf, is a lognormal, defined as:

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma\,x} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] \qquad \text{with } x > 0 . \tag{A.23}$$

Given a sequence of M continuous values $X = \{x_1, x_2, \ldots, x_M\}$, 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:

$$P(x, x + \Delta x) = \int_x^{x+\Delta x} p(y)\, dy . \tag{A.24}$$

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

1. **Collection of rtf:**

   For each concept c, we 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 takes a value between k and $k + \Delta k$, using Equation A.24.

2. **Extraction of fit parameters:**

   Since the form of the lognormal distribution, Equation A.23, the parameters μ and σ that determine it are directly calculated from the empirical rtf list, l_{rtf}, as $\mu \equiv \langle \ln(l_{rtf}) \rangle$ and $\sigma \equiv \sigma\big(\ln(l_{rtf})\big)$, where the last term denotes of the standard deviation of the logarithm of the term-frequency density l_{rtf}.

3. **Computation of the residual entropy:**

   After obtaining parameters μ and σ, we compute the residual entropy, S_d, using a discrete version of the Kullback-Leibler divergence given by:

$$S_d = \sum P(k, k+\Delta k) \ln \frac{P(k, k+\Delta k)}{Q(k, k+\Delta k)} \Delta k, \tag{A.25}$$

   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:

```
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(min(rescaled_tfs)*0.999),\
                      np.log10(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
```

```
27
28 th_pdf = lognorm.pdf( centers_bins_nnz, loc=0, s=sigma, scale=scale )
29
30 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.

Bibliography

[1] Van De Donk W, Loader BD, Nixon PG, Rucht D (2004) *Cyberprotest: New media, citizens and social movements*. (George Routledge & Sons Ltd., London).

[2] González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. *Nature* 453:779–782.

[3] Lew AA, Hall CM, Timothy DJ (2011) *World regional geography: Human mobilities, tourism destinations, sustainable environments*. (Kendall Hunt Publishing Company).

[4] Kuhn T, Perc M, Helbing D (2014) Inheritance patterns in citation networks reveal scientific memes. *Physical Review X* 4(4):041036.

[5] Börner K, Scharnhorst A (2009) Visual conceptualizations and models of science. *Journal of Informetrics* 3(3):161 – 172.

[6] Kuhn T (1962) *The Structure of Scientific Revolutions*. (University of Chicago Press, Chicago).

[7] Wouters PF (1999) Doctoral thesis (University of Amsterdam). Available online at http://garfield.library.upenn.edu/wouters/wouters.pdf.

[8] Bernal JD (1939) *The Social Function of Science*. (George Routledge & Sons Ltd., London).

[9] Garfield E (2009) From the science of science to scientometrics visualizing the history of science with histcite software. *Journal of Informetrics* 3(3):173 – 179.

[10] Börner K, Chen C, Boyack KW (2003) Visualizing knowledge domains. *Annual Review of Information Science and Technology* 37(1):179–255.

[11] Garfield E (1955) Citation indexes for science: A new dimension in documentation through association of ideas. *Science* 122(3159):108–111.

[12] Garfield E (1964) Science citation index - a new dimension in indexing. *Science* 144(3619):649–654.

[13] de Solla Price DJ (1962) *Science since Babylon*. (Yale University Press, New Haven, London).

[14] de Solla Price DJ (1963) *Little science, big science*. (Columbia University Press, New York).

[15] de Solla Price DJ (1965) Networks of Scientific Papers. *Science* 149(3683):510–515.

[16] Cohen MR (1933) Scientific method in *Encyclopaedia of the social sciences*, eds. Seligman ERA, Johnson A. (MacMillan & Co, New York) Vol. 10.

[17] (2017) Microsoft Academic search engine. Available at: https://academic.microsoft.com/.

[18] (2017) Journal Storage digital library. Available at: http://www.jstor.org/.

[19] (2017) ScienceOpen research and publishing network. Available at: https://www.scienceopen.com/.

[20] (2017) Scopus abstract and citation database. Available at: https://www.scopus.com/search/form.uri?display=basic.

[21] (2017) Web of Science citation indexing. Available at: https://clarivate.com/products/web-of-science/.

[22] Martin T, Ball B, Karrer B, Newman MEJ (2013) Coauthorship and citation patterns in the physical review. *Physical Review E* 88(1):012814.

[23] 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.

[24] 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.

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

[26] Organization for Economic Co-operation and Development (2017) Education at a glance 2017, (Paris), Technical report.

[27] Koslowsky RK (2004) *A World Perspective Through 21st Century Eyes: The Impact of Science on Society*. (Trafford Publishing).

[28] Russell B (2016) *The impact of science on society*. (Routledge, New York).

[29] van Raan AFJ (2000) On growth, ageing, and fractal differentiation of science. *Scientometrics* 47(2):347–362.

[30] de Meis L, Leta J (1997) Modern science and the explosion of new knowledge. *Biophysical Chemistry* 68(1):243 – 253.

[31] Jones BF (2009) The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies* 76(1):283–317.

[32] Jones BF (2011) As science evolves, how can science policy? *Innovation policy and the economy* 11:103–131.

[33] Casadevall A, Fang FC (2014) Specialized science. *Infection and Immunity* 82(4):1355–1360.

[34] Shibata N, Kajikawa Y, Takeda Y, Matsushima K (2008) Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation* 28(11):758 – 775.

[35] Bettencourt LM, Kaiser DI, Kaur J (2009) Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics* 3(3):210 – 221.

[36] Prabhakaran T, Lathabai HH, Changat M (2015) Detection of paradigm shifts and emerging fields using scientific network: A case study of information technology for engineering. *Technological Forecasting and Social Change* 91(Supplement C):124 – 145.

[37] Morris SA, Yen GG (2004) Crossmaps: Visualization of overlapping relationships in collections of journal papers. *Proceedings of the National Academy of Sciences of the United States of America* 101(suppl 1):5291–5296.

[38] Newman MEJ (2001) The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America* 98(2):404–409.

[39] Newman MEJ (2004) Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America* 101(Supplement 1):5200–5205.

[40] Börner K, Maru JT, Goldstone RL (2004) The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences of the United States of America* 101(suppl 1):5266–5273.

[41] Milojević S (2014) Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences of the United States of America* 111(11):3984–9.

[42] Petersen AM (2015) Quantifying the impact of weak, strong, and super ties in scientific careers. *Proceedings of the National Academy of Sciences of the United States of America* 112(34):E4671–E4680.

[43] Jones BF, Wuchty S, Uzzi B (2008) Multi-university research teams: Shifting impact, geography, and stratification in science. *Science* 322(5905):1259–1262.

[44] Grauwin S, Jensen P (2011) Mapping scientific institutions. *Scientometrics* 89:943–954.

[45] Gargiulo F, Carletti T (2014) Driving forces of researchers mobility. *Scientific Reports* 4:4860.

[46] Pan RK, Kaski K, Fortunato S (2012) World citation and collaboration networks: uncovering the role of geography in science. *Scientific Reports* 2:902.

[47] Xie Z, et al. (2017) Feature analysis of multidisciplinary scientific collaboration behaviors: A case study on PNAS. *arXiv preprint arXiv:1706.05858*.

[48] Guimerà R, Uzzi B, Spiro J, Amaral LAN (2005) Team assembly mechanisms determine collaboration network structure and team performance. *Science* 308(5722):697–702.

[49] Sun X, Kaur J, Milojević S, Flammini A, Menczer F (2013) Social dynamics of science. *Scientific reports* 3:1069.

[50] Chen P, Xie H, Maslov S, Redner S (2007) Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics* 1(1):8–15.

[51] Wang D, Song C, Barabási AL (2013) Quantifying long-term scientific impact. *Science* 342(6154):127–132.

[52] Ke Q, Ferrara E, Radicchi F, Flammini A (2015) Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences of the United States of America* 112(35):1–6.

[53] Redner, S. (1998) How popular is your paper? an empirical study of the citation distribution. *European Physical Journal B* 4(2):131–134.

[54] Redner S (2005) Citation Statistics from 110 Years of Physical Review. *Physics Today* 58(6):49–54.

[55] Parolo PDB, et al. (2015) Attention decay in science. *Journal of Informetrics* 9(4):734 – 745.

[56] Hirsch JE (2005) An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America* 102(46):16569–16572.

[57] Petersen AM, et al. (2014) Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences of the United States of America* 111(43):15316–15321.

[58] Sinatra R, Wang D, Deville P, Song C, Barabási AL (2016) Quantifying the evolution of individual scientific impact. *Science* 354(6312).

[59] Way SF, Morgan AC, Clauset A, Larremore DB (2017) The misleading narrative of the canonical faculty productivity trajectory. *Proceedings of the National Academy of Sciences of the United States of America* 114(44):E9216–E9223.

[60] Leydesdorff L, Rafols I (2009) A global map of science based on the isi subject categories. *Journal of the American Society for Information Science and Technology* 60(2):348–362.

[61] van Eck NJ, Waltman L (2010) Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics* 84(2):523–538.

[62] (2017) Paperscape interactive map of `arXiv` . Available at: http://paperscape.org/.

[63] Mane KK, Börner K (2004) Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences of the United States of America* 101(Suppl 1):5287–5290.

[64] Boyack KW, Klavans R, Börner K (2005) Mapping the backbone of science. *Scientometrics* 64(3):351–374.

[65] Ahlgren P, Colliander C (2009) Document–document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics* 3(1):49 – 63.

[66] Boyack KW, et al. (2011) Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PloS one* 6(3):e18029.

[67] van Eck NJ, Waltman L, Noyons ECM, Buter RK (2010) Automatic term identification for bibliometric mapping. *Scientometrics* 82(3):581–596.

[68] Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America* 101 Suppl:5228–5235.

[69] Lancichinetti A, et al. (2015) High-Reproducibility and High-Accuracy Method for Automated Topic Classification. *Physical Review X* 5(1):011007.

[70] Silva FN, Amancio DR, Bardosova M, da F. Costa L, Jr. ONO (2016) Using network science and text analytics to produce surveys in a scientific topic. *Journal of Informetrics* 10(2):487–502.

[71] Jurafsky D, Martin J (2000) *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, Prentice Hall series in artificial intelligence. (Prentice Hall, Upper Saddle River, NJ, USA).

[72] Manning CD, Raghavan P, Schütze H (2008) *Introduction to Information Retrieval*. (Cambridge University Press, Cambridge, UK).

[73] De Saussure F (1959) *Course in general linguistics*. (McGraw-Hill, New York, NY, USA).

[74] Waldrop M (1993) *Complexity: The Emerging Science at the Edge of Order and Chaos*, A Touchstone Book. (Simon & Schuster).

[75] Mitchell M (2009) *Complexity: A Guided Tour*. (Oxford University Press).

[76] Jardine N, van Rijsbergen C (1971) The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval* 7(5):217–240.

[77] Voorhees EM (1986) Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing & Management* 22(6):465 – 476.

[78] Hearst MA, Pedersen JO (1996) Reexamining the cluster hypothesis: Scatter/gather on retrieval results in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96. (ACM, New York, NY, USA), pp. 76–84.

[79] Newman MEJ (2010) *Networks: An Introduction*. (Oxford University Press, New York).

[80] Barabási AL (2016) *Network Science*. (Cambridge University Press).

[81] Newman MEJ (2003) The structure and function of complex networks. *SIAM Review* 45(2):167–256.

[82] Albert R, Barabási AL (2002) Statistical mechanics of complex networks. *Reviews of modern physics* 74(1):47.

[83] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. *Physics Reports* 424(4-5):175–308.

[84] Dorogovtsev SN, Mendes JF (2002) Evolution of networks. *Advances in Physics* 51(4):1079–1187.

[85] Milgram S (1967) The small world problem. *Psychology Today* 2:60–67.

[86] Adamic LA, Glance N (2005) The political blogosphere and the 2004 U.S. election: Divided they blog in *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05. (ACM, New York, NY, USA), pp. 36–43.

[87] Bakshy E, Messing S, Adamic LA (2015) Exposure to ideologically diverse news and opinion on Facebook. *Science* 348(6239):1130–1132.

[88] Carro A, Toral R, San Miguel M (2016) Coupled dynamics of node and link states in complex networks: a model for language competition. *New Journal of Physics* 18(11):113056.

[89] Wasserman S, Faust K (1994) *Social Network Analysis: Methods and Applications*. (Cambridge University Press, Cambridge).

[90] Isella L, et al. (2011) What's in a crowd? Analysis of face-to-face behavioral networks. *Journal of theoretical biology* 271(1):166–180.

[91] Pagani GA, Aiello M (2013) The power grid as a complex network: a survey. *Physica A: Statistical Mechanics and its Applications* 392(11):2688–2700.

[92] Pastor-Satorras R, Vespignani A (2007) *Evolution and structure of the Internet: A statistical physics approach*. (Cambridge University Press).

[93] Pastor-Satorras R, Vázquez A, Vespignani A (2001) Dynamical and correlation properties of the Internet. *Physical Review Letters* 87(25):258701.

[94] De Domenico M, Arenas A (2017) Modeling structure and resilience of the Dark Network. *Physical Review E* 95(2):022313.

[95] Guimera R, Mossa S, Turtschi A, Amaral LAN (2005) The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences of the United States of America* 102(22):7794–7799.

[96] Strano E, et al. (2017) The scaling structure of the global road network. *Open Science* 4(10).

[97] Kaluza P, Kölzsch A, Gastner MT, Blasius B (2010) The complex network of global cargo ship movements. *Journal of the Royal Society Interface* 7(48):1093–1103.

[98] Alon U (2006) *An introduction to systems biology: design principles of biological circuits*. (CRC press).

[99] Williams RJ, Martinez ND (2000) Simple rules yield complex food webs. *Nature* 404(6774):180–183.

[100] Chen CY, Ho A, Huang HY, Juan HF, Huang HC (2014) Dissecting the human protein-protein interaction network via phylogenetic decomposition. *Scientific reports* 4:7153.

[101] Buchanan M, Caldarelli G, De Los Rios P, Rao F, Vendruscolo M (2010) *Networks in Cell Biology*. (Cambridge University Press).

[102] Bascompte J, Jordano P, Melián CJ, Olesen JM (2003) The nested assembly of plant–animal mutualistic networks. *Proceedings of the National Academy of Sciences of the United States of America* 100(16):9383–9387.

[103] Bullmore E, Sporns O (2012) The economy of brain network organization. *Nature Reviews Neuroscience* 13(5):336–349.

[104] Cunningham W, Zuev K, Krioukov D (2017) Navigability of the universe. *Scientific Reports* 7:8699.

[105] Coutinho B, et al. (2016) The network behind the cosmic web. *arXiv preprint arXiv:1604.03236*.

[106] Rinaldo A, Rodriguez-Iturbe I, Rigon R, Ijjasz-Vasquez E, Bras RL (1993) Self-organized fractal river networks. *Physical Review Letters* 70(6):822–825.

[107] Anderson PW (1972) More is different. *Science* 177(4047):393–396.

[108] Bondy JA, Murty USR (1976) *Graph theory with applications*. (Macmillan, London) Vol. 290.

[109] Bollobás B (1979) *Graph Theory: An Introductory course*. (Springer-Verlag, New York).

[110] Asratian AS, Denley TM, Häggkvist R (1998) *Bipartite graphs and their applications*. (Cambridge University Press) Vol. 131.

[111] Latapy M, Magnien C, Vecchio ND (2008) Basic notions for the analysis of large two-mode networks. *Social Networks* 30(1):31 – 48.

[112] Cattuto C, Loreto V, Pietronero L (2007) Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences of the United States of America* 104(5):1461–1464.

[113] Liljeros F, Edling CR, Amaral LAN, Stanley HE, Åberg Y (2001) The web of human sexual contacts. *Nature* 411(6840):907–908.

[114] Newman MEJ (2001) Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E* 64(1):016131.

[115] Liu JL, Wang J, Yu ZG, Xie XH (2017) Fractal and multifractal analyses of bipartite networks. *Scientific Reports* 7:45588.

[116] Newman MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. *Physical Review E* 64(2):026118.

[117] Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509–512.

[118] Ramasco JJ, Dorogovtsev SN, Pastor-Satorras R (2004) Self-organization of collaboration networks. *Physical Review E* 70(3):036106.

[119] Zhou T, Ren J, Medo M, Zhang YC (2007) Bipartite network projection and personal recommendation. *Physical Review E* 76(4):046115.

[120] Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America* 101(11):3747–3752.

[121] Newman MEJ (2001) Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E* 64(1):016132.

[122] Padrón B, Nogales M, Traveset A (2011) Alternative approaches of transforming bimodal into unimodal mutualistic networks. the usefulness of preserving weighted information. *Basic and Applied Ecology* 12(8):713–721.

[123] Newman MEJ (2004) Analysis of weighted networks. *Physical Review E* 70(5):056131.

[124] Jørgen Bang-Jensen GZG (2009) *Digraphs: Theory, Algorithms and Applications*. (Springer-Verlag, London).

[125] Breiger RL (1974) The Duality of Persons and Groups. *Social Forces* 53(2):181–190.

[126] Granovetter MS (1973) The strength of weak ties. *American Journal of Sociology* 78(6):1360–1380.

[127] Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393(6684):440–442.

[128] Bollobás B (1985) *Random Graphs*. (Academic Press, London).

[129] Barrat, A., Weigt, M. (2000) On the properties of small-world network models. *The European Physical Journal B* 13(3):547–560.

[130] Travers J, Milgram S (1969) An Experimental Study of the Small World Problem. *Sociometry* 32(4):425–443.

[131] Albert R, Jeong H, Barabási AL (1999) Internet: Diameter of the World-Wide Web. *Nature* 401:130–131.

[132] Montoya JM, Solé RV (2002) Small world patterns in food webs. *Journal of theoretical biology* 214(3):405–412.

[133] Bassett DS, Bullmore E (2006) Small-World Brain Networks. *The Neuroscientist* 12(6):512–523.

[134] Stauffer D, Aharony A (1994) *Introduction to percolation theory*. (Taylor & Francis, London).

[135] Rombach P, Porter MA, Fowler JH, Mucha PJ (2014) Core-periphery structure in networks. *SIAM Journal on Applied Mathematics* 74(1):167–190.

[136] Zhang X, Martin T, Newman MEJ (2015) Identification of core-periphery structure in networks. *Physical Review E* 91(3):032803.

[137] Colizza V, Vespignani A, Serrano MA, Flammini A (2006) Detecting rich-club ordering in complex networks. *Nature Physics* 2(2):110–115.

[138] Sah P, Singh LO, Clauset A, Bansal S (2014) Exploring community structure in biological networks with random graphs. *BMC Bioinformatics* 15(1):220.

[139] Arenas A, Danon L, Díaz-Guilera A, Gleiser PM, Guimerá R (2004) Community analysis in social networks. *The European Physical Journal B* 38(2):373–380.

[140] Fortunato S (2010) Community detection in graphs. *Physics Reports* 486(3):75–174.

[141] Porter M, Onnela JP, Mucha PJ (2009) Communities in networks. *Notices of the American Mathematical Society* 56(9):1082–1097,1164–1166.

[142] Fortunato S, Hric D (2016) Community detection in networks: A user guide. *Physics Reports* 659:1–44.

[143] Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. *Physical Review E* 69(2):026113.

[144] Newman MEJ (2006) Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America* 103:8577–82.

[145] Newman MEJ (2016) Equivalence between modularity optimization and maximum likelihood methods for community detection. *Physical Review E* 94(5):052315.

[146] Fortunato S, Barthélemy M (2007) Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America* 104(1):36–41.

[147] Good BH, de Montjoye YA, Clauset A (2010) Performance of modularity maximization in practical contexts. *Physical Review E* 81(4):046106.

[148] Guimerà R, Sales-Pardo M, Amaral LAN (2004) Modularity from fluctuations in random graphs and complex networks. *Physical Review E* 70(2):025101.

[149] Reichardt J, Bornholdt S (2006) When are networks truly modular? *Physica D: Nonlinear Phenomena* 224(1):20 – 26.

[150] Reichardt J, Bornholdt S (2007) Partitioning and modularity of graphs with arbitrary degree distribution. *Physical Review E* 76(1):015102.

[151] Andrews GE (1976) *The theory of partitions*. (Addison-Wesley, Boston).

[152] Cormen TH, Leiserson CE, Rivest RL (1990) *Introduction to algorithms*. (MIT Press and McGraw-Hill).

[153] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10):P10008.

[154] Yang Z, Algesheimer R, Tessone CJ (2016) A Comparative Analysis of Community Detection Algorithms on Artificial Networks. *Scientific Reports* 6:30750.

[155] Coscia M, Giannotti F, Pedreschi D (2011) A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining* 4(5):512–546.

[156] Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America* 105(4):1118–1123.

[157] Peixoto TP (2014) Efficient monte carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E* 89(1):012804.

[158] Rosvall M, Axelsson D, Bergstrom CT (2009) The map equation. *The European Physical Journal Special Topics* 178(1):13–23.

[159] Rosvall M, Bergstrom CT (2010) Mapping change in large networks. *PloS one* 5(1):e8694.

[160] Yang J, Leskovec J (2014) Structure and overlaps of ground-truth communities in networks. *ACM Trans. Intell. Syst. Technol.* 5(2):26:1–26:35.

[161] Newman MEJ, Peixoto TP (2015) Generalized communities in networks. *Physical Review Letters* 115(8):088701.

[162] Jaccard P (1901) Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles* 37:241–272.

[163] Hric D, Darst RK, Fortunato S (2014) Community detection in networks: Structural communities versus ground truth. *Physical Review E* 90(6):062805.

[164] Levandowsky M, Winter D (1971) Distance between sets. *Nature* 234(5323):34–35.

[165] Kosub S (2016) A note on the triangle inequality for the Jaccard distance. *arXiv preprint arXiv:1612.02696*.

[166] MacKay DJC (2003) *Information Theory, Inference & Learning Algorithms*. (Cambridge University Press, New York, NY, USA).

[167] Danon L, Díaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment* 2005(09):P09008.

[168] Meilă M (2007) Comparing clusterings—an information based distance. *Journal of Multivariate Analysis* 98(5):873 – 895.

[169] He H, Garcia EA (2009) Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng.* 21(9):1263–1284.

[170] Zhang P (2015) Evaluating accuracy of community detection using the relative normalized mutual information. *Journal of Statistical Mechanics: Theory and Experiment* 2015(11):P11006.

[171] Gates AJ, Wood IB, Hetrick WP, Ahn YY (2017) On comparing clusterings: an element-centric framework unifies overlaps and hierarchy. *arXiv preprint arXiv:1706.06136*.

[172] Serrano MÁ, Boguñá M, Vespignani A (2009) Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America* 106(16):6483–8.

[173] Radicchi F, Ramasco JJ, Fortunato S (2011) Information filtering in complex weighted networks. *Physical Review E* 83(4):046101.

[174] Mastrandrea R, Squartini T, Fagiolo G, Garlaschelli D (2014) Enhanced reconstruction of weighted networks from strengths and degrees. *New Journal of Physics* 16(4):043022.

[175] Coscia M, Neffke FMH (2017) Network backboning with noisy data in *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. pp. 425–436.

[176] Squartini T, Garlaschelli D (2011) Analytical maximum-likelihood method to detect patterns in real networks. *New Journal of Physics* 13(8):083001.

[177] Squartini T, Fagiolo G, Garlaschelli D (2011) Randomizing world trade. I. A binary network analysis. *Physical Review E* 84(4):046117.

[178] Squartini T, Fagiolo G, Garlaschelli D (2011) Randomizing world trade. II. A weighted network analysis. *Physical Review E* 84(4):046118.

[179] Dianati N (2016) Unwinding the hairball graph: pruning algorithms for weighted complex networks. *Physical Review E* 93(1):012304.

[180] Gemmetto V, Cardillo A, Garlaschelli D (2017) Irreducible network backbones: unbiased graph filtering via maximum entropy. *arXiv preprint arXiv:1706.00230*.

[181] Zemansky MW, Dittman RH (1997) *Heat and Thermodynamics: An Intermediate Textbook*. (McGraw-Hill, London).

[182] Chandler D (1987) *Introduction to Modern Statistical Mechanics*. (Oxford University Press, New York).

[183] Shannon C (1948) A mathematical theory of communication. *Bell System Technical Journal* 27:379–423.

[184] Borda M (2011) *Fundamentals in information theory and coding*. (Springer).

[185] Kullback S, Leibler RA (1951) On information and sufficiency. *Ann. Math. Statist.* 22(1):79–86.

[186] Kullback S (1959) *Information theory and statistics*. (John Wiley & Sons, New York, NY, USA).

[187] Cover TM, Thomas JA (2012) *Elements of information theory*. (John Wiley & Sons, New York, NY, USA).

[188] Rényi A (1961) On measures of entropy and information in *Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability*. (University of California Press, Berkeley, CA, USA), Vol. 1, pp. 547–561.

[189] Tsallis C (1988) Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics* 52:479–487.

[190] Gell-Mann M, Tsallis C (2004) *Nonextensive entropy: interdisciplinary applications*. (Oxford University Press, Oxford).

[191] Tsallis C (2009) *Introduction to Nonextensive Statistical Mechanics – Approaching a complex world*. (Springer-Verlag, New York) Vol. 34.

[192] Lagrange JL (1811) *Mécanique analytique*. (Ve Courcier, Paris) Vol. 1.

[193] Landau LD, Lifshitz EM (1976) *Mechanics. Vol. 1 (3rd Ed.)*. (Butterworth–Heinemann).

[194] Bertsekas DP (1999) *Nonlinear Programming*. (Athena Scientific, Cambridge).

[195] Vapnyarskii IB (1990) Lagrange multipliers in *Encyclopedia of Mathematics*, ed. Hazewinkel M. (Springer, Dordrecht, The Netherlands) Vol. 5.

[196] Jaynes ET (1957) Information Theory and Statistical Mechanics. *Physical Review* 106(4):620–630.

[197] Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. *SIAM Review* 51(4):661–703.

[198] Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America* 105(45):17268–72.

[199] Evans TS, Hopkins N, Kaube BS (2012) Universality of performance indicators based on citation and reference counts. *Scientometrics* 93(2):473–495.

[200] Goldberg SR, Anthony H, Evans TS (2015) Modelling citation networks. *Scientometrics* 105(3):1577–1604.

[201] Clough JR, Gollings J, Loach TV, Evans TS (2015) Transitive reduction of citation networks. *Journal of Complex Networks* 3(2):189–203.

[202] Garfield E (1994) Scientography: Mapping the tracks of science. *Current Contents: Social & Behavioural Sciences* 7(45):5–10.

[203] Small HG (1978) Cited documents as concept symbols. *Social Studies of Science* 8(3):327–340.

[204] Bradford SC (1934) Sources of information on specific subjects. *Engineering: An Illustrated Weekly Journal* 137:85–86.

[205] Bianconi G (2015) Interdisciplinary and physics challenges of network theory. *EPL (Europhysics Letters)* 111(5):56001.

[206] Gibney E (2014) How to tame the flood of literature. *Nature* 513(7516):129–130.

[207] Ahlgren P, Jarneving B (2008) Bibliographic coupling, common abstract stems and clustering: A comparison of two document-document similarity approaches in the context of science mapping. *Scientometrics* 76(2):273–290.

[208] Manning CD, Schütze H (1999) *Foundations of statistical natural language processing*. (MIT Press, Cambridge, USA) Vol. 999.

[209] Shi F, Chen L, Han J, Childs P (2017) A data-driven text mining and semantic network analysis for design information retrieval. *Journal of Mechanical Design* 139(11):111402.

[210] Dhohoon Kim L, Jang DH (2017) Expert views on innovation and bureaucratization of science: Semantic network analysis of discourses on scientific governance. *Science and Public Policy* 45(1):36–44.

[211] Vilhena D, et al. (2014) Finding Cultural Holes: How Structure and Culture Diverge in Networks of Scholarly Communication. *Sociological Science* 1(15):221–238.

[212] Chen C (2006) CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology* 57(3):359–377.

[213] Aberer K, Boyarsky A, Cudré-Mauroux P, Demartini G, Ruchayskiy O (2011) ScienceWISE: a Web-based Interactive Semantic platform for scientific collaboration in *10th International Semantic Web Conference (ISWC 2011 - Demonstration Track)*.

[214] Aberer K, Boyarsky A, Cudré-Mauroux P, Demartini G, Ruchayskiy O (2011) ScienceWISE: a Web-based Interactive Semantic platform for scientific collaboration in *10th International Semantic Web Conference (ISWC 2011 - Outrageous Ideas)*.

[215] Boyarsky A, et al. (2013) From scientific papers to the scientific ontology: dynamical clustering of heterogeneous graphs and ontology crowdsourcing in *Joint Workshop on Large and Heterogeneous Data and Quantitative Formalization in the Semantic Web (LHD + SemQuant 2012)*.

[216] Prokofyev R, et al. (2012) Tag recommendation for large-scale ontology-based information systems in *11th International Semantic Web Conference (ISWC 2012 - Evaluations and Experiments Track)*.

[217] Astafiev A, Prokofyev R, Guéret C, Boyarsky A, Ruchayskiy O (2012) ScienceWISE: A Web-based Interactive Semantic platform for paper annotation and ontology editing in *Extended Semantic Web Conference (ESWC 2012 - Satellite Events)*.

[218] Prokofyev R, Demartini G, Cudré-Mauroux P, Boyarsky A, Ruchayskiy O (2013) Ontology-based word sense disambiguation in the scientific domain in *35th European Conference on Information Retrieval (ECIR 2013)*.

[219] Constantin A (2014) Doctoral thesis (The University of Manchester). Available online at https://www.escholar.manchester.ac.uk/uk-ac-man-scw:230124.

[220] Spärck Jones K (1972) A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(1):11–21.

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

[222] Martini A, et al. (2016) Sciencewise: Topic modeling over scientific literature networks. *arXiv preprint arXiv:1612.07636*.

[223] Martini A, Cardillo A, De Los Rios P (2017) Entropic selection of concepts in networks of similarity between documents. *arXiv preprint arXiv:1705.06510*.

[224] Ferrer i Cancho R, Solé R (2003) Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America* 100:788–791.

[225] Font-Clos F, Boleda G, Corral Á (2013) A scaling law beyond Zipf's law and its relation to Heaps' law. *New Journal of Physics* 15(9):093033.

[226] Visser M (2013) Zipf's law, power laws and maximum entropy. *New Journal of Physics* 15(4):043021.

[227] Gerlach M, Altmann EG (2014) Scaling laws and fluctuations in the statistics of word frequencies. *New Journal of Physics* 16(11):113010.

[228] Yan X, Minnhagen P (2015) Maximum Entropy, Word-Frequency, Chinese Characters, and Multiple Meanings. *PloS one* 10(5):e0125592.

[229] Zipf GK (1949) *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology*. (Addison-Wesley).

[230] A. Berger and S. D. Pietra and V. D. Pietra (1996) A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics* 22(1):39–71.

[231] Hotho A, Nürnberger A, Paaß G (2005) A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology* 20:19–62.

[232] Baek SK, Bernhardsson S, Minnhagen P (2011) Zipf's law unzipped. *New Journal of Physics* 13(4):043004.

[233] Herrera JP, Pury PA (2008) Statistical keyword detection in literary corpora. *The European Physical Journal B* 63(1):135–146.

[234] Carpena P, Bernaola-Galván PA, Carretero-Campos C, 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. *Physical Review E* 94(5):052302.

[235] Altmann EG, Dias L, Gerlach M (2017) Generalized entropies and the similarity of texts. *Journal of Statistical Mechanics: Theory and Experiment* 2017(1):014002.

[236] (2017) Interactive Sankey diagrams. Available at: http://www.bifi.es/~cardillo/data.html#semantic. A standalone version can be downloaded at https://www.dropbox.com/sh/tpmth8xk1hgk39b/AABWi6SNfSYCG2xNQz_L1ZFva?dl=0.

[237] Tufte ER (1983) *The Visual Display of Quantitative Information*, Encyclopedia of mathematics and its applications. (Graphics Press, Cheshire, CT, USA).

[238] Palchykov V, Gemmetto V, Boyarsky A, Garlaschelli D (2016) Ground truth? Concept-based communities versus the external classification of physics manuscripts. *EPJ Data Science* 5(1):28.

[239] (2017) The Twitter Developer Documentation, API overview. Available at: https://dev.twitter.com/overview/api.

[240] Blei DM, Lafferty JD (2009) Topic models in *Text mining: classification, clustering, and applications*, Data Mining and Knowledge Discovery Series, eds. Srivastava A, Sahami M. (Chapman & Hall/CRC, Boca Raton, FL, USA), pp. 71–94.

[241] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391–407.

[242] Deerwester S (1988) Improving Information Retrieval with Latent Semantic Indexing in *Proceedings of the 51st ASIS Annual Meeting (ASIS '88)*, eds. Borgman CL, Pai EYH. (American Society for Information Science, Atlanta, GA, USA), Vol. 25, pp. 36–40.

[243] Landauer TK, Dumais ST (1997) A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104(2):211–240.

[244] JC (1990) The singular-value decomposition and its use to solve least-squares problems in *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation*. (Adam Hilger, Bistol, UK), 2 edition, pp. 30–48.

[245] Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401:788–791.

[246] Xu W, Liu X, Gong Y (2003) Document clustering based on non-negative matrix factorization in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '03. (ACM, New York, NY, USA), pp. 267–273.

[247] Pauca VP, Shahnaz F, Berry MW, Plemmons RJ (2004) Text mining using non-negative matrix factorizations in *Proceedings of the 2004 SIAM International Conference on Data Mining*, SIAM proceedings series. (SIAM, Philadelphia, PA, USA), pp. 452–456.

[248] Landauer TK, McNamara DS, Dennis S, Walter K (2007) *Handbook of Latent Semantic Analysis*. (Psychology Press, New York, NY, USA).

[249] Hofmann T (1999) Probabilistic latent semantic analysis in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. (Morgan Kaufmann Publishers Inc.), pp. 289–296.

[250] Hofmann T (2001) Unsupervised Learning by probabilistic Latent Semantic Analysis. *Machine Learning* 42(1):177–196.

[251] Ding C, Li T, Peng W (2008) On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis* 52(8):3913 – 3927.

[252] Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. *Journal of Machine learning Research* 3:993–1022.

[253] Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM (2009) Reading tea leaves: How humans interpret topic models in *Advances in neural information processing systems*. pp. 288–296.

[254] Yau CK, Porter A, Newman N, Suominen A (2014) Clustering scientific documents with topic modeling. *Scientometrics* 100(3):767–786.

[255] Liu L, Tang L, Dong W, Yao S, Zhou W (2016) An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* 5(1):1608.

[256] Jelodar H, Wang Y, Yuan C, Feng X (2017) Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *arXiv preprint arXiv:1711.04305*.

[257] Liu L, Tang L, He L, Zhou W, Yao S (2016) An overview of hierarchical topic modeling in *8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*. Vol. 1, pp. 391–394.

[258] Jensen S, Liu X, Yu Y, Milojevic S (2016) Generation of topic evolution trees from heterogeneous bibliographic networks. *Journal of Informetrics* 10(2):606 – 621.

[259] Gerlach M, Peixoto TP, Altmann EG (2017) A network approach to topic models. *ArXiv preprint arXiv:1706.05858*.

[260] Blei DM, Lafferty JD (2007) A correlated topic model of *Science*. *Annals of Applied Statistics* 1(1):17–35.

[261] Knight WR (1966) A computer method for calculating Kendall's tau with ungrouped data. *Journal of the American Statistical Association* 61(314):436–439.

[262] (2017) Physics Subject Headings (PhySH). Available at: https://physh.aps.org/.

[263] Newman MEJ, Clauset A (2016) Structure and inference in annotated networks. *Nature Communications* 7.

[264] Peel L, Larremore DB, Clauset A (2017) The ground truth about metadata and community detection in networks. *Science Advances* 3(5).

[265] Sachdev S (2011) Condensed Matter and AdS/CFT in *From Gravity to Thermal Gauge Theories: the AdS/CFT Correspondence*. (Springer, Berlin, Heidelberg), pp. 273–311.

[266] Pires AST (2014) *AdS/CFT Correspondence in Condensed Matter*, 2053-2571. (Morgan & Claypool Publishers).

[267] (2016) Scientific buzzwords obscure meaning. *Nature* 538(7624):140.

[268] Scudellari M (2017) Big science has a buzzword problem. *Nature* 541(7638):450–453.

[269] Shibata N, Kajikawa Y, Matsushima K (2007) Topological analysis of citation networks to discover the future core articles. *Journal of the American Society for Information Science and Technology* 58(6):872–882.

[270] (2016) `arXiv` submission rate statistics. https://arxiv.org/help/stats/2016_by_area/index.

[271] Warner SN (2001) arXiv, the OAI and peer review (Workshop on the Open Archives Initiative (OAI) and Peer Review journals in Europe (OAI1)). Available at: https://indico.cern.ch/event/408493/contributions/1858146/attachments/818845/1123315/expanded_talk.pdf.

[272] Planck collaboration (2014) Planck 2013 results. I. Overview of products and scientific results. *Astronomy & Astrophysics* 571:A1.

[273] Arkani-Hamed N, Finkbeiner DP, Slatyer TR, Weiner N (2009) A theory of dark matter. *Physical Review D* 79(1):015014.

[274] Bertone G, Hooper D, Silk J (2005) Particle dark matter: evidence, candidates and constraints. *Physics Reports* 405(5):279 – 390.

[275] Spiro M (1995) Dark matter. *Nuclear Physics B - Proceedings Supplements* 43(1):100 – 107.

[276] Fornengo N (2008) Status and perspectives of indirect and direct dark matter searches. *Advances in Space Research* 41(12):2010–2018.

[277] Leskovec J, Rajaraman A, Ullman JD (2014) *Mining of massive datasets*. (Cambridge University Press, Cambridge).

[278] Sorzano COS, Vargas J, Montano AP (2014) A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*.

[279] Berndt DJ, Clifford J (1994) Using dynamic time warping to find patterns in time series in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94. (AAAI Press), Vol. 10, pp. 359–370.

[280] Harris TE (1963) *The theory of branching processes*. (Springer-Verlag, Berlin).

[281] Nguyen HC, Zecchina R, Berg J (2017) Inverse statistical problems: from the inverse ising problem to data science. *Advances in Physics* 66(3):197–261.

[282] (2017) The Medical Subject Headings of the United States Library of Medicine. Available at: https://www.nlm.nih.gov/mesh/.

[283] Olesen JM, Bascompte J, Dupont YL, Jordano P (2007) The modularity of pollination networks. *Proceedings of the National Academy of Sciences of the United States of America* 104(50):19891–19896.

[284] (2017) Web of Science search tool. Available at: https://webofknowledge.com/.

[285] (2017) Python software foundation. python language reference (version 2.7). Available at: http://www.python.org.

[286] Jones E, Oliphant T, Peterson P, , Plenz D (2001) Scipy: Open source scientific tools for python. Available at: https://www.scipy.org.

[287] Johansson, F *et al.* (2013) mpmath: a Python library for arbitrary-precision floating-point arithmetic (version 0.18). Available at: http://mpmath.org/.

[288] Alstott J, Bullmore E, Plenz D (2010) Powerlaw: a Python package for analysis of heavy-tailed distributions. *PloS one* 9(1):e85777.

Andrea Martini

Curriculum Vitae

Professional address EPFL SB IPHYS LBS
BSP 519 (Cubotron UNIL)
Route de la Sorge
CH-1015 Lausanne
Switzerland

E-mail andrea.martini@epfl.ch

Telephone +41 78 726 55 60

Date of birth July 21th, 1988
Citizenship Italian
Gender Male

Actual position

November 1st, 2013 - present **PhD student in Physics**
Laboratoire de Biophysique Statistique, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
Supervisor: Prof. Paolo De Los Rios

Education

2010 - 2013 **Master degree in Physics of Complex Systems**
University of Turin, Italy
Degree mark: 110/110 *cum laude*

Master thesis
Activity patterns and community structure of time-varying human interaction networks
Institute for Scientific Interchange (ISI) Foundation, Turin, Italy
Supervisor: Dr. Ciro Cattuto

2011 - 2012 Exchange period within ERASMUS program
University of Granada, Spain

2007 - 2010 **Bachelor degree in Physics**
University of Turin, Italy
Degree mark: 108/110

Thesis
Theoretical and experimental characterization of an heralded single-photon source
National Institute of Metrological Research (INRiM), Turin, Italy
Supervisor: Dr. Marco Genovese

2002 - 2007 **High school**
Liceo Scientifico Marie Curie, Pinerolo, Italy

Internships

July - August 2010 *Study of the properties of a low-noise heralded single-photon source based on parametric down-conversion*,
National Institute of Metrological Research (INRiM), Turin, Italy

Research interests

Statistical physics
Complex systems
Complex networks
Data mining

Publications

2017 *ScienceWISE: Topic Modeling over Scientific Literature Networks*
preprint paper @ arXiv : http://arxiv.org/abs/1612.07636.

2017 *Entropic selection of concepts in networks of similarity between documents*
preprint paper @ arXiv : https://arxiv.org/abs/1705.06510. (Under review)

Conferences and Schools

September 16 - 19, 2013 IceLab Camp - Communication and creativity camp
Vindeln, Sweden

April 7 - 18, 2014 Complex Networks Thematic School
Les Houches, France

June 1 - 5, 2015 NetSci - International School and Conference on Network Science
Zaragoza, Spain

September 3 - 8, 2015 Mediterranean School of Complex Networks
Salina, Sicily

March 23 - 25, 2016 CompleNet 2016 – 7th Workshop on Complex Networks
Dijon, France

July 11 - 13, 2016 Complex Networks: from theory to interdisciplinary applications – satellite meeting of the StatPhys16
Marseilles, France

September 19 - 22, 2016 CCS2016 – Conference on Complex Systems 2016
Amsterdam, The Netherlands

Teaching

Feb. 2017 - May 2017
Feb. 2016 - May 2016 Teaching assistant at exercise sessions of Analytical Mechanics,
Bachelor degree in Physics, EPFL

Sept. 2016 - Dec. 2016 Teaching assistant at exercise sessions of Statistical Physics of Biomacromolecules,
Master degree in Physics, EPFL

Sept. 2015 - Dec. 2015 Teaching assistant at exercise sessions of Statistical Physics III,
Master degree in Physics, EPFL

Feb. 2015 - May 2015
Feb. 2014 - May 2014 Teaching assistant at exercise sessions of Physical Biology of the Cell II,
Bachelor degree in Life Sciences and Technologies, EPFL

Sept. 2014 - Dec 2014 Teaching assistant at programming sessions of Computational Physics I,
Bachelor degree in Physics, EPFL

Awards

2013 Scholarship promoted by "Fondazione Franco e Marisa Caligara" aimed at rewarding a Master thesis project about interdisciplinary knowledge

Other activities

Winter 2012 - Spring 2013 Tutor of high school students in performing Physics laboratory experiments, as part of the Italian school orientation program "Piano Nazionale per le Lauree Scientifiche"

Information technology skills

Operating systems GNU/Linux, Windows
Programming languages Python, Matlab, bash, Mathematica, Netlogo, C++
Markup LaTeX
Programs Microsoft Word, Excel, Powerpoint

Language skills

Italian Native
English General understanding, speaking and writing: very good
Spanish General understanding, speaking and writing: very good
Certificate of course in Spanish as a foreign language, CLM Granada;
Level B2.2

French General understanding, speaking and writing: good

References

Prof. Paolo De Los Rios Laboratoire de Biophysique Statistique, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
e-mail: paolo.delosrios@epfl.ch

Dr. Ciro Cattuto Institute for Scientific Interchange (ISI) Foundation, Turin, Italy
e-mail: ciro.cattuto@isi.it