QUANTIFYING CONVERGENCE IN THE SCIENCES

SARA LUMBRERAS, *Universidad de Comillas;*
PENNY MEALY, *Institute of Economic Thinking, University of Oxford*
CHRISTOPHER VERZIJL, *ABN AMRO Private Banking International;*
SAMUEL F. WAY, *University of Colorado Boulder*

ABSTRACT: Traditional epistemological models classify knowledge into separate disciplines with different objects of study and specific techniques, with some frameworks even proposing hierarchies (such as Comte's). According to thinkers such as John Holland or Teilhard de Chardin, the advancement of science involves the convergence of disciplines. This proposed convergence can be studied in a number of ways, such as how works impact research outside a specific area (citation networks) or how authors collaborate with other researchers in different fields (collaboration networks). While these studies are delivering significant new insights, they cannot easily show the convergence of different topics within a body of knowledge. This paper attempts to address this question in a quantitative manner, searching for evidence that supports the idea of convergence in the content of the sciences themselves (that is, whether the sciences are dealing with increasingly the same topics). We use Latent Dirichlet Analysis (LDA), a technique that is able to analyze texts and estimate the relative contributions of the topics that were used to generate them. We apply this tool to the corpus of the Santa Fe Institute (SFI) working papers, which spans research on Complexity Science from 1989 to 2015. We then analyze the relatedness of the different research areas, the rise and demise of these sub-disciplines over time and, more broadly, the convergence of the research body as a whole. Combining the topic structure obtained from the collected publication history of the SFI community with techniques to infer hierarchy and clustering, we reconstruct a picture of a dynamic community which experiences trends, periodically recurring topics, and shifts in the closeness of scholarship over time. We find that there is support for convergence, and that the application of quantitative methods such as LDA to the study of knowledge can provide valuable insights that can help researchers navigate the increasingly wide literature as well as identifying potentially fruitful areas for research collaboration.

KEY WORDS: convergence, topic modelling, latent dirichlet allocation, complex adaptive systems.

¿Convergen las diferentes disciplinas de conocimiento? evidencia cuantitativa

RESUMEN: Los modelos epistemológicos tradicionales clasifican el conocimiento en disciplinas separadas con objetos de estudio distintos y técnicas específicas, incluso proponiendo esquemas jerárquicos (por ejemplo, Comte). Según pensadores como John Holland o Teilhard de Chardin, el avance de la ciencia implica una convergencia entre sus disciplinas. Esta convergencia puede estudiarse de maneras distintas, como el impacto de diferentes autores fuera de su equipo (redes de citación) o la manera en la que colaboran (redes de coautoría). Aunque estos estudios están generando ideas interesantes, no son capaces de mostrar la convergencia de los distintos temas que se tratan en un cuerpo de trabajos. Este artículo intenta estudiar esta pregunta desde un punto de vista cuantitativo, buscando evidencias que apoyen la idea de convergencia en el contenido de las ciencias en sí mismas (es decir, si las ciencias se ocupan de temas cada vez más cercanos entre ellos). Empleamos Latent Dirichlet Analysis (LDA), una técnica que analiza textos y estima las contribuciones relativas de los temas que los generan (estos temas se definen como distribuciones de palabras). Aplicamos esta técnica al corpus de artículos publicados por el Instituto de Santa Fe (Santa Fe Institute, SFI), que describe trabajos relacionados con las Ciencias de la Complejidad entre 1989 y 2015. Analizamos la cercanía entre las diferentes áreas, la aparición y desaparición de temas de investigación y, en general, la posible convergencia entre disciplinas. Combinando la estructura obtenida de la historia de las publicaciones de SFI con técnicas de inferencia de jerarquía y clustering, reconstruimos la perspectiva de una comunidad científica dinámica que experimenta tendencias, temas recurrentes y cambios en la cercanía de las diferentes disciplinas. Nuestros resultados muestran que hay evidencias

PENSAMIENTO, vol. 71 (2015), núm. 269, pp. 1383-1399
DOI: pen.v71.i269.y2015.020

de convergencia y que la aplicación de métodos cuantitativos puede proporcionar nuevos elementos de comprensión que ayuden a los investigadores a estructurar una literatura científica cada vez más amplia y compleja, así como a identificar áreas potenciales para nuevas colaboraciones.

PALABRAS CLAVE: convergencia, modelado de temas, latent dirichlet allocation, sistemas complejos adaptativos.

1. Introduction

Traditional epistemological models classify knowledge in separate disciplines with different objects of study and specific techniques. Some of these classifications even assign hierarchical levels to each branch of knowledge, with Comte's Theory of Science providing one of the most well recognized frameworks (Comte, 1868). For Comte, all known disciplines could be arranged into a continuous from mathematics to astronomy, physics, chemistry, biology and, lastly, sociology. The order in which Comte ordered the sciences reflected increasing complexity and generality of the subject, which contrasted with a decreasing complexity of the instruments applied by each discipline.

However, the thesis of the unity of science has long been recognized. For example, Oppenheim proposed the unity of science as a working hypothesis (Oppenheim & Putnam, 1958) and von Bertalanffy set out a more holistic framework for unifying natural and social sciences in his General Systems Theory (Bertalanffy, 1968).

From this viewpoint, the multiplicity of disciplines provides different perspectives and frameworks for understanding and interpreting observed phenomena. However, the advancement of science then necessarily entails some convergence of its different fields. Such convergence has been proposed by, among others, John Holland and Teilhard de Chardin. John Holland was one of the main contributors to the field of complexity research, and, in particular, the definition of complex adaptive systems and their properties (Holland, 2012). A Complex Adaptive System (CAS) is defined as a system that has a large numbers of components, often called «agents», which interact and adapt or learn. A typical CAS has properties such as self-similarity, complexity, emergence and self-organization. Science as a whole can be viewed as a CAS, where the different disciplines evolve together to create a common, emerging holistic view.

This idea of convergence in the scientific disciplines also resonates with the ideas of Teilhard de Chardin (De Chardin & Wall, 1965); knowledge first starts in the divergence of the plurality of disciplines, but they should then start to converge into the final, single truth. Henry Kenny, who was himself a Teilhardian scholar, described the convergence of several scientific disciplines into one synthesis, evolution, in the following words: «Other evidence for evolution from comparative anatomy, genetics, physiology, biochemistry, biogeography, taxonomy and embryology, together with the currently available preponderance of paleontological evidence constitutes an evidential convergence that literally renders any other explanation besides evolution as almost unthinkable». The latter can be understood as one example of convergence in the sciences, with topic evolution bringing together comparative anatomy, genetics, physiology or biochemistry.

Further examples of this kind of phenomenon were given, more recently, by E.O. Wilson and his *consilience* theory (Wilson & Ros, 1999). Consilience is defined as an agreement between the approaches to a topic of different academic subjects, especially in science and the humanities. Consilience is therefore a particular case of convergence, where the truth about one topic is revealed through a multiplicity of paths. We (the

authors) understand convergence in the context of knowledge in a more general way. All topics are subjected to consilience as the scope of each of the sciences grows to also encompass the others, and extends the methods it uses to incorporate those that have been proposed and applied in the other disciplines.

E.O. Wilson identifies in his work many recent examples of consilience, such as the unification of Darwin's evolution with genetics, the unification of forces in modern physics and the unification of Brownian motion and atomic theory. Although attractive and powerful, this evidence is qualitative in nature. Our research proposal is to study the hypothesis of convergence in the sciences in a quantitative manner.

Trends and patterns within scientific knowledge have been analyzed quantitatively previously. For instance, citation networks, which represent the citation of one document by another, have been used to understand how new knowledge builds on existing literature (Leicht, Clarkson, Shedden, & Newman, 2007; Rice, Borgman, & Reeves, 1988). Similarly, the study of co-authorship networks, which depict the joint work or collaboration between authors, has revealed a number of habits and collaboration patterns within academic communities (Grossman, 2002; Ioannidis, 2008; Newman, 2001; Newman, 2001; Newman, 2001). Rather than examining knowledge convergence in terms of «who is citing who», or «who is working with who», we instead seek to examine the underlying knowledge *content* directly. Using an approach known as *topic modelling*, we investigate how particular topics arise and evolve in terms of their thematic content. Specifically, we seek to measure the convergence of topics across disciplines, that is, to what extent different disciplines are referring to the same topics and using the same concepts.

Topic modeling is a statistical technique for identifying particular topics in a corpus of documents (Blei, Ng, & Jordan, 2003; Blei, 2012). Intuitively, the method is based on two key premises. Firstly, abstract topics can be represented in terms of a set of words that frequently appear together. For example, a topic like «gardening», might be represented in terms of words like «flower», «soil», and «seed» while a topic like «cricket» might be represented in terms of words like «bat», «ball», and «runs». Secondly, if a document discusses a particular topic, one might expect it to contain more of certain words and less of others. In general, documents are likely to encompass multiple topics. One of the advantages of topic modeling is that it uses the relative frequency of topic words in the document to provide a quantitative estimate of the extent to which a document focuses on one topic relative to another. For example, if a document primarily discusses gardening but also talks about cricket, we would expect the document to contain many more «gardening words» than «cricket words».

In order to study the convergence of knowledge across different sub-disciplines, we focus on an inherently interdisciplinary research area, Complexity Science. Unlike most research disciplines, Complexity Science is not based on a single theory or framework. Rather, it encompasses theories from a wide variety of disciplines and employs a number of approaches to study the aforementioned properties of complex adaptive systems (Mitchell, 1992). While science is one exemplification of a complex adaptive system, there are many other diverse and wide-reaching examples, such as economies (in which economic agents interact in markets), ecosystems (in which species interact with each other and their environment), traffic (in which drivers respond to the position and speed of other drivers on the road) and the internet, (in which humans interact in cyberspace with both humans and computers). The breadth in the fields contributing to Complexity Science (including computer science, biology, mathematics, and physics),

and the diversity of its applications, makes the research area a particularly interesting body of knowledge to examine.

To analyze key patterns and knowledge convergence within the Complexity Science literature, we focus on research output from the Santa Fe Institute. The Santa Fe Institute is a research organization dedicated to furthering research in Complexity Science and has been instrumental to the ongoing development of the field since the mid-1980s. In this paper, we study the entire corpus of Santa Fe Institute working papers, which comprise approximately 1600 online research documents published from 1989 to mid-2015.

By analyzing the frequency of words within each document and across the corpus as a whole, we quantitatively identify 20 topics corresponding to key subject areas represented in the literature. We first examine each of the 20 topics in terms of the overlap or proximity in knowledge content relative to all the other topics. We find that some topics are quite close to each other, while other topics comprise content that is quite different. We then investigate the popularity of individual topics (how often they are used) within the literature and how popularity changes over time. Finally, we quantify the proximity between documents at different time periods and find evidence for knowledge convergence within the corpus.

2. Topic Modelling using Latent Dirichlet Allocation

To extract the thematic content underlying the documents in the SFI corpus, we applied Latent Dirichlet Allocation (LDA), an unsupervised, generative framework to discover latent «topics» or groupings of semantically related words (Blei et al., 2003). Within this framework, each topic defines a unique probability distribution (a multinomial) over all words in a vocabulary, and each document contains words drawn from a mixture or probability distribution over these topics. The generative story for producing documents under LDA is as follows:

For each document,

1. Draw a distribution over topics
2. For each word being added to the document,
    a. Select a topic from the document's distribution over topics
    b. Select a word, drawn from the selected topic (a distribution over words)

The «latent» aspect of LDA stems from the fact that we do not know *a priori* what the topic distributions are for each document or what the topics themselves are. These elements are expressed as latent variables within the model and are inferred from data, using statistical inference techniques.

To perform LDA, we used MALLET, a topic modeling toolkit, which includes implementations of the LDA algorithm as well as several other popular topic models and utilities for preprocessing text files (McCallum, 2002). Prior to running MALLET, we preprocessed the corpus using standard approaches from natural language processing. In particular, we applied lemmatization (converting words to their respective lemmas — e.g. «converting» to «convert») and removed «stop words», words that appear frequently in the English language but provide little information for determining the thematic content of a document — e.g. «a», «the».

In the following sections, we use LDA to analyze the SFI corpus in to two ways. First, we analyze the topics themselves to determine what concepts are discussed in this literature. Second, we view the corpus as a time series and examine how the topics' popularity changes over time. This analysis corresponds to Holland's concept of CAS flows. In the following sections we describe and present the results of several preliminary analyses.

3. Static Topic Proximity Dendrogram

In our study of the SFI corpus, we used LDA to obtain document-topic probability distributions, «topic distributions» hereafter, which describe the thematic content underlying each document. As a first analysis, we investigated the similarities between the topics themselves by quantifying similarities in their usage patterns. When applied to the SFI corpus, LDA identifies the following topics (i.e. distributions of words). The table below shows these topics and their most frequent words.

	Topic	***Most used words***							
1	Molecular Biology	structure	sequence	landscape	gamma	space	secondary	base	neutral
2	Quantitative Finance	market	price	order	trader	good	stock	money	equilibrium
3	Particle Physics	energy	system	phys	function	spin	temperature	entropy	state
4	Dynamical Systems	model	system	dynamic	equation	time	point	parameter	state
5	Genetics	gene	protein	sequence	genome	expression	interaction	evolution	acid
6	Immunology	cell	model	antibody	antigen	immune	rate	clone	affinity
7	Information Theory	state	process	information	entropy	machine	complexity	causal	measure
8	Mathematics	function	theorem	proof	case	space	matrix	lemma	problem
9	Game Theory	agent	game	strategy	player	equilibrium	action	model	utility
10	Cellular Automata	gamma	rule	automaton	state	lattice	particle	site	figure
11	Complex Systems	system	theory	language	science	complex	structure	process	object
12	Early Civilisations	population	area	site	society	patch	resource	university	press
13	Political Theory	social	company	system	market	network	party	control	political
14	Statistics	data	model	distribution	time	result	series	size	number
15	Genetic Algorithms	fitness	population	genetic	mutation	landscape	selection	function	generation
16	Networks	network	graph	node	number	degree	vertex	random	edge
17	Chemistry	specie	reaction	extinction	food	evolution	model	diversity	rate
18	Social Norms and Cooperation	group	individual	social	behavior	cost	model	level	member
19	Technology, Cities and Growth	firm	technology	city	economic	cost	production	economy	income
20	Learning Algorithms	algorithm	problem	system	learning	computer	function	input	time

Table 1. *TOPICS and their most used words*

We used Jensen-Shannon Divergence (JSD, discussed further in the appendix) to compare topics' usage across the corpus. JSD is used extensively in probability theory and in natural language processing as a measure of similarity between two probability distributions. For our purposes, JSD allows us to quantify the relatedness of two topics by comparing how their usage is distributed throughout the corpus. The idea is that if,

across all documents, two given topics are always allocated a similar relative weight then they must be related (when one appears, the other appears as well with a certain dominance). That is, intuitively, we understand two topics to be related when the documents that deal with one topic also tend to deal with the other.

We applied this to the SFI corpus and visualized the analysis in a dendrogram, shown in Fig. 1. The dendrogram shows the hierarchical clustering of topics (Ward Jr, 1963), implemented in Matlab, where topics that are closer merge earlier in the diagram, and topics that are further apart merge later. This visual representation is useful to understand the relationships among the topics in an easily interpretable way.

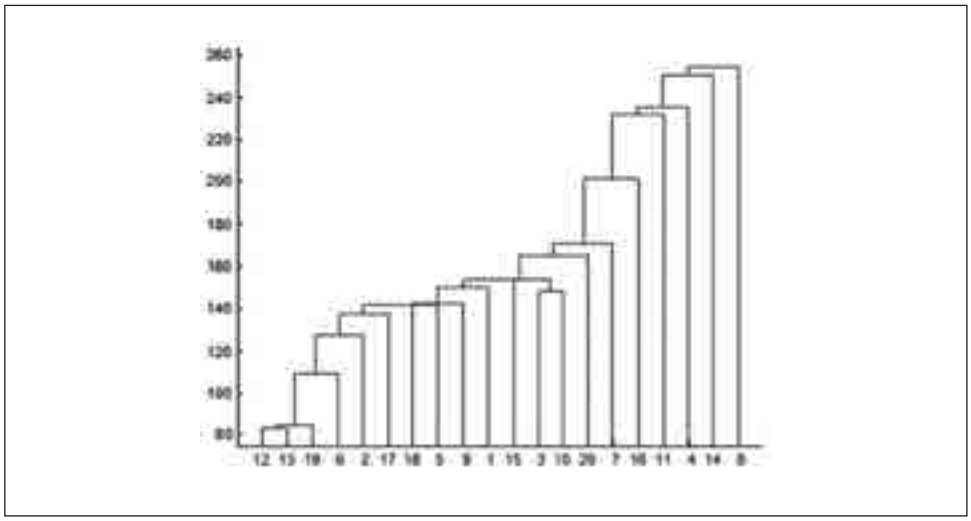

FIGURE 1. *Topic Proximity Dendrogram*

As can be seen, the topics that appear closer in the corpus and so merge the earliest in the dendrogram are Early Civilizations (12), Political Theory (13) and Technology, Cities and Growth (19). In this case we have three topics in the social sciences that relate to different forms of societal organization.

If we continue analyzing the dendrogram, we arrive at the merge of topics that deal with biology and economy, with genetic algorithms being responsible for this merge: Biochemistry (17), Social Norms and Cooperation (18), Genetics (5), Game Theory (9), Molecular Biology (1) and Genetic Algorithms (15).

At the extreme of the spectrum, we can find the topics Statistics (14) and Mathematics (8), which are the latest to be incorporated to the dendrogram. This means that, within the range of topics, these two represent subjects that tend to appear in an isolated way or together with a wide variety of other topics so that no specific connections can be made. This would be typical of topics that can be considered representative of tools (in this case, mathematical tools). These topics would appear in papers that carry out theoretical developments or that apply them to a range of other problems with not one single application being dominant over the others.

4. Topic Dynamics

In addition to analyzing the proximity between topics, we also investigated how individual topics differ in relative prevalence (or popularity) and how this changes over time. Our measure of relative prevalence is derived from the normalized topic probability distributions across each document. By construction, each document will have a non-negative probability across all 20 topics. However, topics that are more strongly represented will have a higher probability value than topics that are less represented in the document.

The five panels below show the normalized topic probability distributions across all documents in the SFI corpus. We have ordered each document in chronological order according to its date. While the documents in the SFI corpus span the time period from 1989 to 2015, documents are not uniformly distributed across time. Consequently, 'time' represented along the x-axis in Figure 2 and Figure 3 is not a uniform distribution of time, but a uniform distribution of chronologically ordered publications. The y-axis categories in each of the panels in Figure 2 correspond to each of the 20 topics analyzed in the corpus and the z-axis represents the topic distribution value.

We have ordered topic categories along the y-axis according to their relative prevalence over the *entire* corpus period. That is, the first few topics listed (such as Statistics and Complex Systems) are topics that consistently have high topic probability values in the majority of documents over time. In contrast, the last topics listed (Early Civilizations and Immunology) only tend to have high topic probability values over a few documents in the corpus.

In considering why some topics might be more prevalent than others, it is helpful to examine some of the words within each topic. The Statistics topic consists of words such as 'data', 'model', 'distribution', 'estimate', 'probability' and 'analysis'. The Complex Systems topic comprises words such as 'theory', 'complex', 'process', 'world', 'evolution' and 'nature'. As these words tend to be of quite a *general* nature, it is not surprising that the topics tend to be more highly represented in a large number of documents. In contrast, the Early Civilizations topic consists of words such as 'habitat', 'settlement', 'social', 'household', 'resource' and 'population', while the Immunology topic comprises words including 'antibody', 'infection, 'immune', 'virus' and 'tumor'. These words are likely to be used in much more specific contexts, which sheds light on why these topics are represented much more sparsely within the corpus.

The five panels of Fig. 2 also give an indication of how topics change in relative prevalence over the corpus time-span. While the relative prevalence of some topics (like Statistics, Complex Systems and Dynamical Systems) remains reasonably constant over time, other topics are shown to experience a marked change.

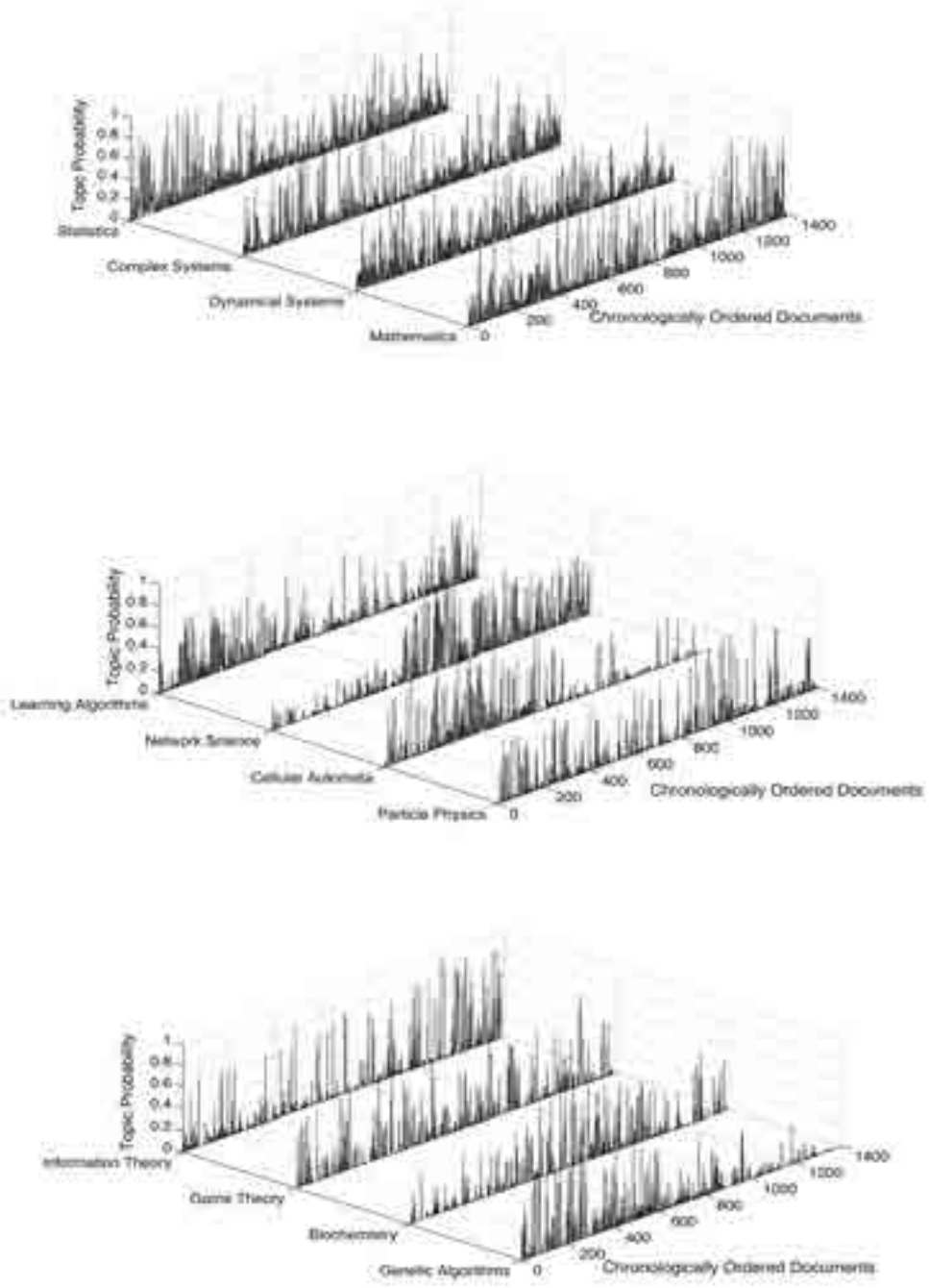

FIGURE 2. *Topic dynamics for top-20 topics in the corpus*

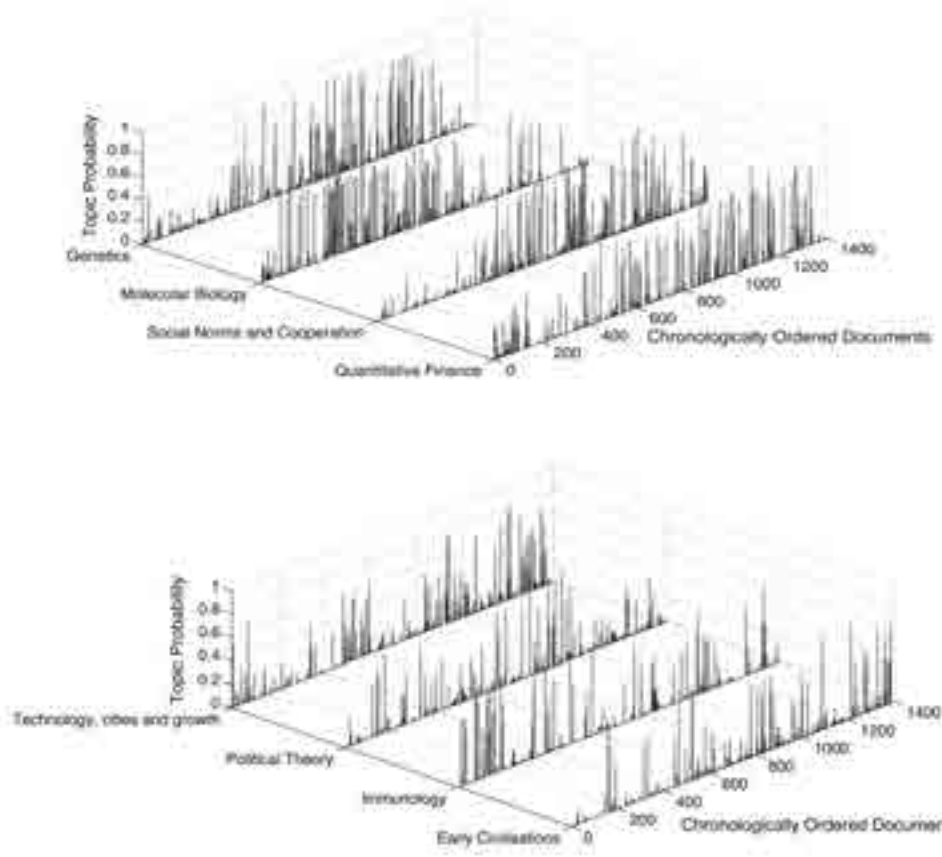

FIGURE 2. *Continued*

In Fig. 3, we show two distinct examples of the 'rise' and 'demise' of topic prevalence. The first panel shows the relative prevalence of the Network Science topic over the corpus time-span. A substantial increase in topic prevalence can be seen after the first 600 chronologically ordered documents, which corresponds to the year 1999. This striking increase in topic prevalence around this time is not surprising, as it corresponds to the period just after Watts and Strogatz (Watts & Strogatz, 1998) published their work on the 'small-world' network and sparked a more mathematical analysis of social networks. With new interest in small world networks rekindled, the year 1999 saw a number of important papers published, such as Barabási and Albert's papers (Barabasi & Albert, 1999) on scaling properties in many real world networks like the world-wide web. Since that period, network science has evidently remained a significantly prevalent topic within the SFI corpus.

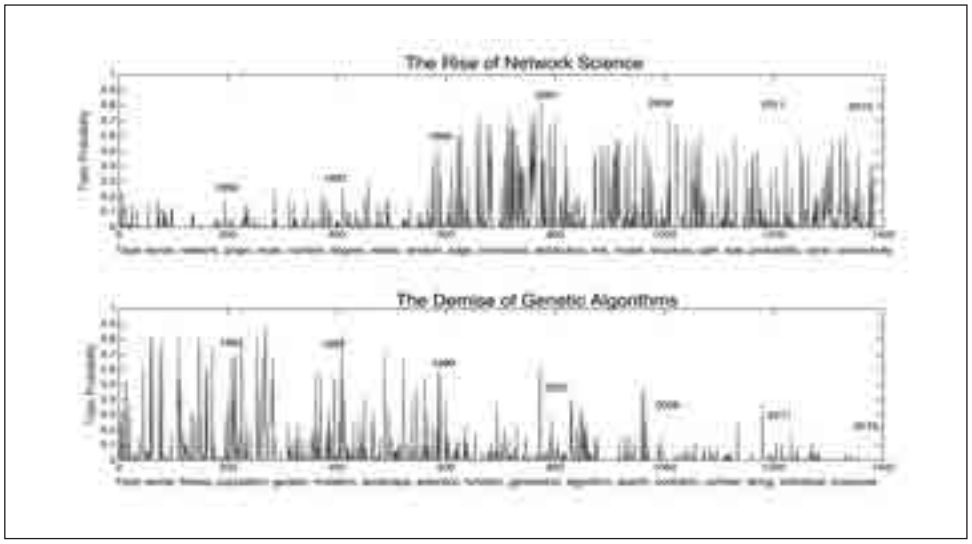

FIGURE 3. *Relative trends in genetic algorithms and network science.*

The second panel in Fig. 3 shows quite a different trend. While the genetic algorithms topic was relatively prevalent within the SFI corpus documents in earlier years (spanning the years 1989-1999), the turn of the century appears to correspond to a decrease in interest in this topic. On the one hand, this trend may be interpreted as being somewhat surprising, particularly in light of SFI's strong historical involvement with Holland's development of genetic algorithms in the 1970s. On the other hand, it might indicate a potential tendency for SFI research to be targeted towards relatively *new* ideas, techniques and research areas. Genetic algorithms were considered 'frontier' research during the 1970's-1990's, particularly as improvements in computers expanded both their power and potential applications. However, with new research areas like statistical machine learning inevitably «crowding out» researcher time devoted to existing topics, older research areas may become less prevalent in present day publications.

5. DOCUMENT RELATIONSHIPS AS HEATMAPS

In preceding sections we discussed the topic hierarchy derived from an LDA analysis by clustering topics using the Jensen-Shannon divergence as a measure of distance. We now apply the same techniques to the structure of the corpus as a whole, over time, by studying the relationships among documents. This is essentially an implementation of the CAS flows concept noted previously. To this end, we again take the topic distributions obtained by LDA (Blei et al., 2003) as implemented in Mallet (McCallum, 2002) for each document and likewise calculate Jensen-Shannon divergences (Endres & Schindelin, 2003; Kullback & Leibler, 1951; Lin, 1991) as a measure of the pairwise distances. We store these in a distance matrix , which we use in two steps of subsequent analysis.

First, we visualize the distance-based hierarchy reconstructed using UPGMA (Unweighted-Pair Group Method with Arithmetic Mean, one of the simplest and most

widely used metrics (Sokal, 1958)) as implemented in scipy.cluster.hierarchy in the SciPy library (Jones, Oliphant, & Peterson, 2001).[1] A result of this, with color-coded clusters is shown in Fig. 4, illustrating a part of the clustering of all documents in the 2010-2015 time window. To better understand of the resulting hierarchies, we also algorithmically label the documents in the tree by creating a string from an ordered list of the top 20 words in the top topic as the label for each document, which can be compared to the titles and abstracts of the respective papers. Second, to investigate the structure of the publications by the SFI community and its evolution over time, this clustering is used to permute the columns and rows of the distance matrix, which we then visualize as a heatmap of distances.

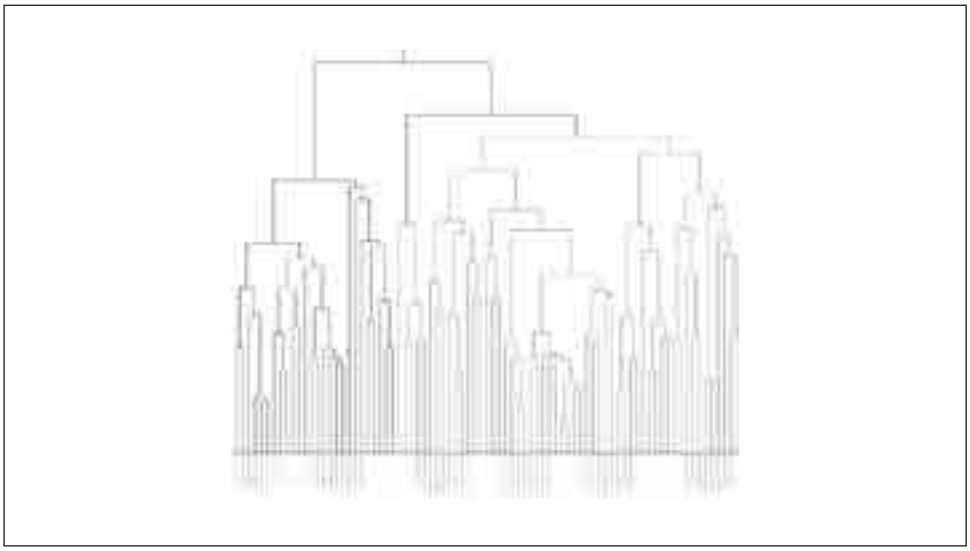

FIGURE 4. *Partial dendrogram of the 2010-2015 block of documents in the corpus showing a cluster of documents with mathematical terms related to networks (black), documents related to neural networks and machine learning (green) and a larger block corresponding to information theory and theory of computation (red).*

The two can be conveniently shown together in a clustered heatmap visualization, which makes the inferred hierarchy explicit, as in Fig. 5, shown for the full corpus («sfi_doc»). We also characterize the distribution by a histogram of the JSD measure, which is centered at 0.55 (scale from 0 to ln 2), with a spread of 0.10 and a heavy tail characterized by excess kurtosis =2.79.

[1] This is a reasonable fallback clustering approach, but the content of the analysis is not too sensitive to the algorithm used, even when clustering is done by other means, such as e.g. Ward's algorithm (Ward Jr, 1963).

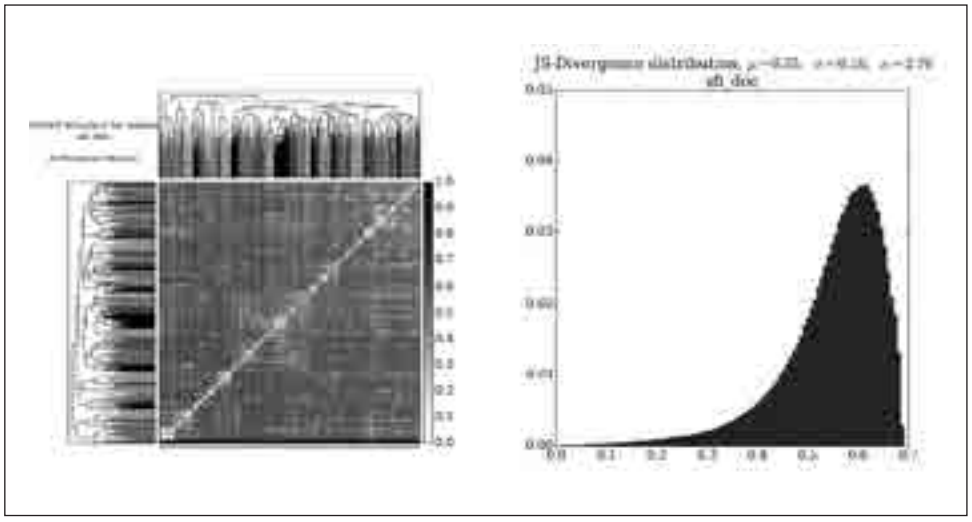

FIGURE 5. *The clustered heatmap shows the large-scale structure of the full corpus, with some banding visible and some block-structure on the diagonal corresponding to the cores of clusters, and the Histogram of the JSD distribution with sample statistics for the full («sfi_doc») corpus.*

The ordering of the documents is again a rough proxy for a time-axis, and so to study changes over time we bin the documents into 5-year groups. An initial impression of the time-dependence can be obtained by creating groups with independent topic models (one for each group), but a more structured result is found by training a single topic model on the full corpus and using it to infer topic and word distributions for each group. Following this procedure, we have tried k=20, 50 and 100 topics, finding 50 to be a reasonable balance between high-level summary topics and very fine-grained detail.[2]

Despite being a fairly coarse-grained method with respect to time, applying this technique to the SFI corpus leads to two key observations. First, that we see structure, and its evolution of over time. Second that this directly reflects the changing focus of the community over time discussed in the preceding section. Note that the rough approach could straightforwardly be improved by using more rigorous techniques explicitly modeling time-dependence (Blei & Lafferty, 2006; Blei, Wang, & Heckerman, 2008).

The results are illustrated in Fig. 6 for three 5-year groups of documents, and suggest a community in change. The overall color-intensity speaks to the average distance between topics, which appears to slowly decrease over time. However, this decrease seems to be driven not by a homogenous reduction in distance, but by the appearance of low-distance block structure (larger, lighter blocks) in the final 2010-2015 block, indicating that a group of more closely linked research appears. In contrast, what we do not find are strong indications of are isolated sub-communities, which would show up as (side-) bands which are internally close (a light block-diagonal square) and otherwise well-separated from the rest of the community (dark bands in relation to documents outside the block).

[2] With 50 topics we obtain 46 clusters using the default clustering threshold of linkages below 0.7 max(linkage)

This visualization of the qualitative trends suggests further exploration by estimating the distance distribution as a function of time-blocks in Fig. 7. This also yields a clearer picture of deviations from the averaged behavior of Fig 8, as the groups contain unequal numbers of documents.

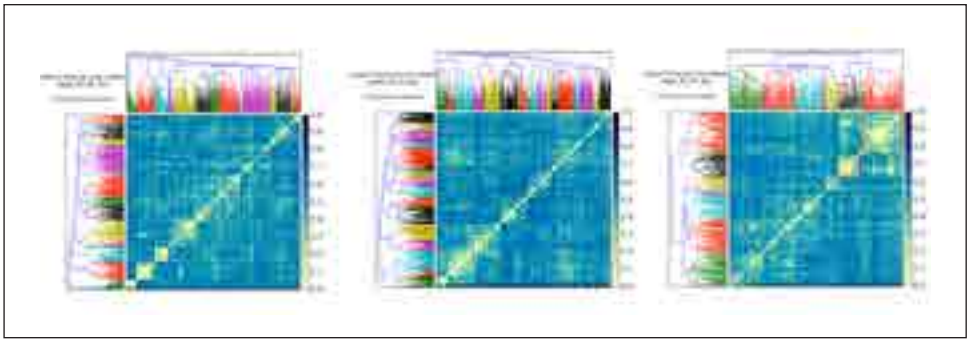

FIGURE 6. *Clustered heatmaps for different blocks, showing shift to lower average distance and emergence of more block structure in the later cluster.*

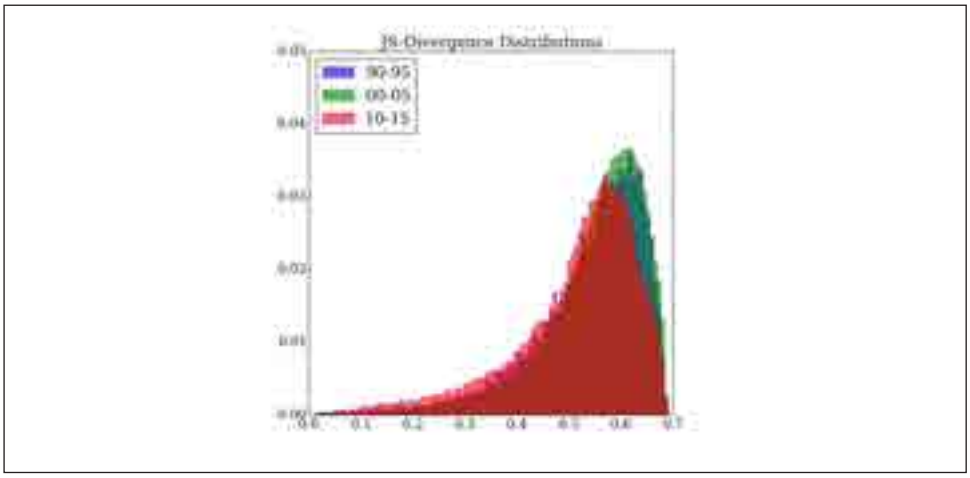

FIGURE 7. *Distributions from Fig. 6 over time, showing a slight broadening and shift to lower mean for the final 5-year block, and increased clustering in the small-distance tail (lower excess kurtosis).*

Overall, we find that the blocks, like the full corpus, have a negatively-skewed distribution peaked at the high-end of the distance distribution, and a spread () of roughly 0.11. However, the mean has been shifting very slowly downward over time (also relative to the document-weighted average). To illustrate the clustering at low distances we use not the skew,[3] but the excess kurtosis . This is a measure of the 'heaviness' of the

[3] Skew, the 3rd moment, is biased by the upper bound of the JSD.

tail, where the decrease in with time corresponds to a heavier tail: i.e. increasingly many documents characterized by small pairwise distances, as the bottom pane

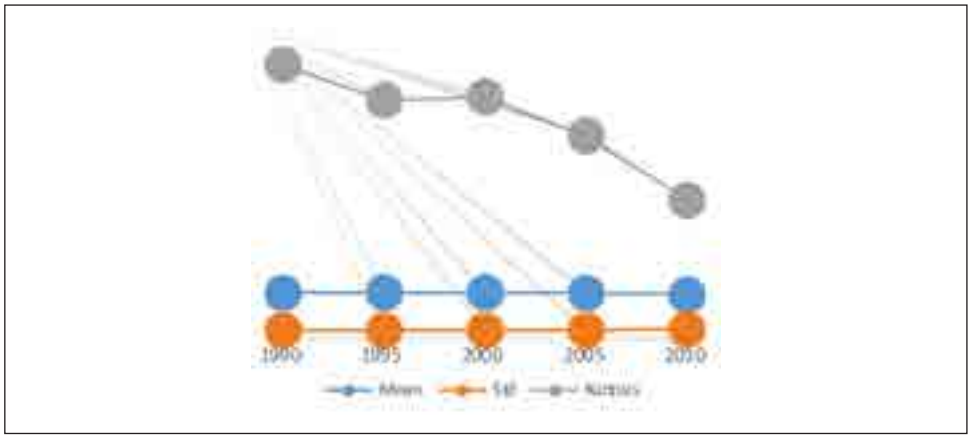

FIGURE 8. *Trends in topic distributions over time.*

The clusters associated with this development (highlighted in Figure 7) correspond to mathematical terms related to networks (black), documents related to neural networks and machine learning (green) and a larger block corresponding to information theory and theory of computation (red), also noted in Fig. 4.

The ease with which we can make this dynamic visually intuitive suggests that these techniques are useful for developing insight into the community structure underlying developments in complexity sciences research at SFI, but also into other communities of knowledge. In faster-moving fields, more rigorous treatment of a consistent topic hierarchy that evolves over time (Blei et al., 2010), together with modeling the time-dependence of the topics (Blei & Lafferty, 2006) should help to significantly strengthen the initial exploration discussed here.

6. DISCUSSION

In this research, we attempted to quantify whether there is convergence in the contents of the sciences, with different disciplines dealing with increasingly the same topics. The results of our analysis are relevant for the study of knowledge on two levels, addressing knowledge generation within a particular community, and presenting a method for the quantified study of hypotheses about the underlying process by which this occurs more broadly.

First, we focus specifically on the community of research on complexity science embodied by the corpus of working papers from the Santa Fe Institute. We have found, by means of a scalable analytic technique, that the picture of a dynamic interdisciplinary community emerges, in which we see evidence both for shifting focus (e.g. away from early emphases on genetic algorithms towards applications and techniques like network science which draw together multiple techniques), and for some measure of convergence in the research. The latter is evidenced by the clustering of documents in the fat tail of

the distance distributions, which reflect a period in which more closely-linked work is being undertaken than in other periods. We remark, however, that this does not yet reflect a convergence of fields. Rather, it suggests the convergence of work within fields, as reflected by the topics of focus in the community and the words used to express them.

Nonetheless, as a trend is evident it would be interesting to revisit the community periodically over longer periods in the future. Seeing continuations of such dynamics would certainly lend support the convergence hypothesis, as well as potentially clarifying details of its mechanism.

Next, we consider our approach as being a useful analytical in its own right, providing a new approach to studying the structure of topics in the sciences. While our methodology is complementary to classification by experts and the study of citation- and co-authorship networks, it has the advantage that it lets the texts «speak for themselves», and allows us a view of the development of communities of knowledge that is intimately connected to their use of language. The techniques, moreover, lead to quantifiable insights via an analysis that is scalable (requiring only limited computer time, relative to time required to read and analyze the texts in full), but also wholly complementary to the study and understanding of experts within a field (in which sense it is only semi-automated, e.g. leaving the interpretation of meaning of topics to those performing such analysis).

Our approach also enables us to identify trends and infer hierarchies of topics based on their prevalence and relative distance to each other. The analysis of these hierarchies provides intuition about how fields of knowledge are organized, where convergence has occurred, and where it may occur in the future. It is important to note that these hierarchies reflect relative relationships expressed by the text, rather than as a strict epistemological statement in the style of Comte's framework. They afford us a novel view on what is going on in a research community, and what the structure of knowledge looks like at a point in time, *represented* as a hierarchy.

Although this paper has primarily focused on providing a means to quantify thinking about convergence as a measure for progress, both in epistemology and the history of science, it is important to stress that our approach could have much broader practical applications in two key areas: improving research efficiency and informing research strategy. Insofar as we have presented a methodology that allows a researcher to automate a large amount of the time and effort that usually needs to be invested in 'taking stock' of the literature, our approach could be applied more generally to drastically improve the efficiency of research. Indeed, with global scientific output estimated to be doubling every nine years (Bornmann & Mutz, 2015), there is a clear need for new tools and techniques to help researchers navigate the growing wilderness of academic literature.

In relation to research strategy, more research institutions are now recognizing the new insights and significant advancements emanating from the cross-pollination of different disciplines. For this reason, collaboration between researchers of different fields is becoming more common. However, as the coming together of ideas is a complex and dynamic process whose effects are scattered throughout the literature, it is often difficult to know *which* research areas of disciplines would be the most fruitful to combine. Here, our approach could be particularly useful, as we have demonstrated a quantitative methodology for understanding which bodies of knowledge are on the rise (in terms of research popularity) and for identifying research areas that are significantly related in terms of knowledge content. Given the scalability and flexibility of our presented techniques, it is likely such tools could be adopted to inform policies governing resource allocation and guide research strategies of both academic institutions and individuals.

7. Acknowledgments

The authors gratefully acknowledge fruitful interaction with the staff and faculty of the Santa Fe Institute, as well as interesting discussions with fellow participants in the 2015 Complex Systems Summer School, in particular with J. Thomas and A. Andreyevna.

8. References

Barabasi, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science (New York, N.Y.), 286*(5439), 509-512.

Baran, M. E., & Wu, F. F. (1989). Optimal capacitor placement on radial distribution systems. *Power Delivery, IEEE Transactions on, 4*(1; formulated and a solution algorithm is proposed. The location, type, and size of capacitors, voltage constraints, and load variations are considered. The objective of capacitor placement is peak power and energy loss reduction, taking into ac(TRUNCADO)), 725-734.

Bertalanffy, L. v. (1968). *General system theory: Foundations, development, applications* Braziller. New York.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM), 57*(2), 7.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning,* pp. 113-120.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993-1022.

Blei, D., Wang, C., & Heckerman, D. (2008). Continuous time dynamic topic models. *Intl. Conference on Machine Learning (ICML),*

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology,*

Cherven, K. (2013). *Network graph analysis and visualization with gephi* Packt Publishing Ltd.

Comte, A. (1868). *The positive philosophy of auguste comte* W. Gowans.

De Chardin, P. T., & Wall, B. (1965). *The phenomenon of man* Harper & Row New York, NY, USA:.

Endres, D. M., & Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory,*

Grossman, J. W. (2002). The evolution of the mathematical research collaboration graph. *Congressus Numerantium,* , 201-212.

Holland, J. H. (2012). *Signals and boundaries: Building blocks for complex adaptive systems* Mit Press.

Ioannidis, J. P. (2008). Measuring co-authorship and networking-adjusted scientific impact. *PloS One, 3*(7), e2778.

Jones, E., Oliphant, T., & Peterson, P. (2001). Others. SciPy: Open source scientific tools for python. *Web Http://www.Scipy.Org,*

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics,* , 79-86.

Leicht, E. A., Clarkson, G., Shedden, K., & Newman, M. E. (2007). Large-scale structure of time evolving citation networks. *The European Physical Journal B, 59*(1), 75-83.

Lin, J. (1991). Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on, 37*(1), 145-151.

McCallum, A. K. (2002). {MALLET: A machine learning for language toolkit}.

Mitchell, W. M. (1992). Complexity: The emerging science at the edge of order and chaos. *Touchstone, New York,*

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences, 98*(2), 404-409.

Oppenheim, P., & Putnam, H. (1958). *Unity of science as a working hypothesis* na.

Rice, R. E., Borgman, C. L., & Reeves, B. (1988). Citation networks of communication journals, 1977-1985 cliques and positions, citations made and citations received. *Human Communication Research, 15*(2), 256-283.

Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull, 38*, 1409-1438.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236-244.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440-442.

Wilson, E. O., & Ros, J. (1999). *Consilience* Circulo de Lectores.

9. APPENDIX

The JSD, which we use as a distance measure for topic distributions, is calculated for a set of probability distributions $M+1(A)$, where A is a set with is endowed with a -algebra. The JSD is thus a function that relates a pair of probability distributions to a positive real value

$$M+1(A)\times M+1(A)\rightarrow[0,\infty),$$

although not yet a metric (D. M. Blei, Griffiths, & Jordan, 2010; Cherven, 2013; Kullback & Leibler, 1951; Lin, 1991) .

JSD is a symmetric and smoothed version of the relative entropy or Kullback-Leibler divergence (P||Q), which is defined as the expectation of the logarithmic difference between the discrete probabilities P(i) and Q(i),

$$D_{KL}(P||Q)=\Sigma_i P(i)\ln[P(i)/Q(i)].$$

JSD is now defined as

$$JSD(P||Q)=1/2\cdot D(P||M)+1/2\cdot D(Q||M), \text{ with}$$
$$M=1/2\cdot(P+Q).$$

JSD gives a measure of distance between two probability distributions. In order to calculate distances between the topics, we define the probability distributions $P_t(d)$, as the normalized transpose of the topic allocation matrix. That is, we define vectors that describe, for each topic, the allocation that each document presents to this topic. We then scale this vector in order to ensure that the total allocation sums 1. These probability distributions have the form $P_t(d)$, where t represents the topics and d the documents in the corpus. We define the distances between topics as the divergences between these distributions, calculated as the JSD.

slumbreras@comillas.edu

Sara Lumbreras,
Penny Mealy,
Christopher Verzijl,
Samuel F. Way

[Artículo aprobado para publicación en diciembre de 2014].