Vive la République! The relationship between science, civil society, and government

Daniel Hopp[1*], Robin Cowan[1,2], Moritz Müller[2], Fabiana Visentin[1]

[1]UNU-Merit, Maastricht University, Maastricht & 6211, Netherlands.

[2]BETA, University of Strasbourg, Strasbourg & 67081, France.

[*]Corresponding author. Email: daniel.hopp@un.org

The relationship between science and other segments of society, such as government or civil society, is a complex and evolving one. We study how the three domains interact. Using a novel dataset of scientific abstracts, government documents, and media articles from 2015-2024, we quantify shifts in topic focus and inter-domain overlap. We find a notable divergence: government and media topics are more closely aligned with each other than either is with science, suggesting that science deals with many topics that, today, are not frequently discussed by the other two domains. Further, we observe that topics rarely transition from being dominated by science to widespread adoption in media and government. Instead, a more common pattern emerges where media acts as an initial amplifier, particularly for technology and digitization-related subjects, subsequently drawing increased attention from government and science. Finally, our analysis highlights media's more reactive and transient attention compared to the slower-changing, more persistent, focus of government and science. These insights collectively underscore the value of creating enduring channels for scientific advice to effectively navigate complex societal challenges and of fundamental research as the foundation for many of the topics that diffuse from media to science and government.

Acknowledgements

The authors declare that this research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. No external funding or research support was received for this work. All authors have read and approved the final manuscript and confirm that there are no conflicts of interest to disclose.

1 Introduction

The relationship between science and other segments of society, such as government or civil society, is a complex and evolving one. In the US, universities first began formalizing their approaches to research in the late 19th and early 20th centuries, with little government involvement. World War 2 and the Cold War Era profoundly realigned the incentive structures and funding models of academia, however. Geopolitical and military considerations injected millions of dollars of funding into fields relevant to them and the government recognized the importance of supporting the scientific community more broadly (*1, 2*). As a result, federal funding came to dominate research budgets in the US. This era conformed well to Michael Polanyi's "Republic of Science" framework for understanding the incentives and culture of scientific research. In Polanyi's view, research agendas should be set intrinsically by the research community, free from centralized authority or decree (*3*). While the central government provided a large proportion of funding, it did not set specific research agendas (*2*). This changed in the 1980s, however, as the percentage of R&D expenditure by the private sector surpassed that of the federal government for the first time since the end of World War 2 (*4*). This change affected research, leading to concerns that the research system was prioritizing short-term results and practical uses instead of long-term goals and expanding general knowledge (*4*). These developments have led many to question the implicit social contract inherent in academic research models. Critics cite a lack of understanding and transparency on science's relevance and value to society and the economy (*5*). Distrust and skepticism towards higher education from the general public reached an all-time high in 2023 (*6*), while global dynamics move funding increasingly away from basic research towards other domains, such as defense (*7*).

This complex and, at times, troubled social relationship with scientific research will only be put under further stress in the coming decades, as societies deal with emerging and worsening economic, environmental, geopolitical, and health challenges. The COVID-19 pandemic was a stark example of the importance of the complex interplay between not only science and government policy, but between science and civil society. Looking ahead to an uncertain future, it is important to quantify, characterize, and understand these dynamics. This research proposes a novel method for quantifying the domains' relationship to each other using topic modelling. South Africa was

chosen as the case study for the paper due to a mixture of practical and substantive considerations. During the apartheid era, academic research was called upon (and in some cases volunteered) to provide scientific underpinnings for the separation of 'races'. Since 1994, of course, this has been abandoned, and the science system (and universities more generally) is called upon to help move the country beyond this, providing knowledge and human capital that will aid social and economic transformation. The rapid expansion of the system, partly aimed at extra-scientific goals, makes it an interesting case. Practically, South Africa's scientific, government, and news media output is published largely in English, which simplifies issues of natural language processing (NLP) and translation. Its volume is also considerably smaller than larger countries such as the United States, which streamlines data collection.

The rest of the paper proceeds as follows: Section 2 provides further background on the development of science's social contract and reviews approaches in the existing literature for probing the domains' relationship to each other. Section 3 details the dataset used for the analysis. Section 4 explains the modelling approach. Section 5 discusses and analyzes the outputs of the model, while section 6 concludes and outlines avenues for future research.

2 Background

2.0.1 The social contract for science

While the post-WWII era saw governments generously funding science with relative autonomy for researchers, largely adhering to Polanyi's "Republic of Science" paradigm, this dynamic has shifted significantly. Since the 1980s, private R&D funding has outstripped government sources, coinciding with growing claims of science's politicization (*4, 8*). Science now faces increased pressure to demonstrate clear social benefits and applications, while governments are urged to better ground policies in research (*4, 9*). This has intensified competition for scarcer, more conditional public funding, leading to a greater focus on short-term, applied projects over basic "blue skies" research (*4*). Consequently, new approaches to measuring the societal impact of research have developed as funding agencies seek ways to allocate limited resources amidst these competing demands (*10*).

Additionally, since the late 1990s, public funding agencies have begun explicitly incorporating social impact considerations into their peer-review process of grant applications (*11*). This has

been a polemic change in the funding allocation process, sparking debates around the fundamental purpose of science as well as the qualifications of scientists in assessing social importance (*12*). Studies and surveys have found that many scientists dislike or find unimportant this increasing emphasis on social impact and find the degree of government involvement in setting research priorities excessive (*11, 12*).

Despite scientists' purported dislike of focus on and conditions of impact, Figure 1 shows it seems to have had a real effect on, if not the content of their papers, at least how they talk about them. Relative usage of the terms 'impact' and 'impacts' nearly quadrupled between 1995 and 2024 in South African paper abstracts.

Further, Myers (*13*) analyzed U.S. National Institute of Health data on Requests for Application (RFA) to see to what degree centralized decision-making can "steer research to certain topics", finding an RFA premium of 65%. In other words, I would need to receive an RFA grant of $165,000 to pursue RFA-stipulated research, while I would accept $100,000 to pursue my own research. Switching costs may explain a high proportion of this premium, with Myers finding RFA applications to be 25% more dissimilar to scientists' previous work compared to open applications. This highlights that dictating research agendas according to policy priorities and social impact may be costly.

In South Africa, research spending has grown significantly since the 1970s, though it represents only 0.8% of GDP, compared to an OECD average of 2.4% (*14*). A large portion of this growth was driven by increases in funding for the Agricultural Research Council (ARC). The ARC, however, has declined as a proportion of South African research funding since the 2010s, with a higher proportion going to universities (*14*). In contrast with the United States, South Africa has experienced a decline in the business sector's proportion of research funding, falling from 56% of all R&D in 2001, to only 39% in 2015 (*14*). This increases South African science's sensitivity to changes in government research funding policy. South Africa is no exception to the international trend of an increasing focus on social impact in research, with government departments involved in research funding openly questioning its impact on South African society and priority issues facing it, such as HIV and the environment (*9*). This focus is borne out in the data, with South African expenditure on applied research growing from 40% of R&D in 2001 to 48% in 2015 (*14*).

Regardless of researchers' views on the matter, the shift has created new difficulties with actually

assessing and evaluating social impact, both ex ante and ex post facto. In the case of the former, who is qualified to evaluate proposals' social impact? How should they make these decisions? Based on economic potential? Impact on human well-being? Environmental impact? In the case of the latter, how can we measure the social impact of existing scientific work? Science's impact on science has a well-established paradigm for its measurement: citations. However, society's output cannot be distilled into homogeneously formatted 20-30 page articles. This led to the emergence of so-called "altmetrics" in the 2010s, which attempt to capture the impact of science beyond traditional paper citation (*15*). Altmetrics incorporate additional measures of impact such as social media mentions, downloads/reads, mentions in press articles, or citations in government documents (*16*).

Traditional citation-based impact measurement as well as altmetrics both attempt to quantify and trace the impacts of specific research outputs. While altmetrics captures effects more broadly than citations alone, it still suffers from an inability to deal with non-countable effects. The research performed in this paper describes and quantifies science's relationship to society and government at a macro level. The focus is rather on creating high-level "mappings" of collective dialog and thought within the three domains, illustrating their spatial and temporal relationships and identifying factors which influence or change those relationships.

In order to convert those mappings into metrics of impact, methods of identifying emergent topics and themes in a body of literature are required. These methods were drawn from work on topic and knowledge discovery and information and semantic science evolution. For instance, the Data Study Group of the Alan Turing Institute (*17*) uses the BERTopic modelling technique to identify emerging trends in scientific literature. Tshitoyan et al. (*18*) used the Word2Vec approach to identify potential future discoveries in materials science. Dias et al. (*19*) produced an analysis of the temporal evolution of linguistic similarity between research domains.

However, most of these studies focus on academia in isolation, not examining its relationship with other domains. Those works that do explicitly focus on society-science or policy-science interactions are either strictly theoretical/discussion papers, such as Leydesdorff (*20*) or Smith (*21*), or focus on citations as the measure of impact, like Bornmann et al. (*22*).

3 Data

Data on scientific research comes from the Scopus abstract and citation database. The database contains peer-reviewed articles and conference papers from more than 7,000 publishers, covering more than 330 disciplines. Scopus was chosen over other alternatives, such as Web of Science, due to its comprehensive coverage of research areas. Actual data collected consists of the abstracts of 449,611 papers published by at least one author with a South African institutional affiliation over the years 1995-2024. Figure 2 shows the number of abstracts collected by year.

Data on general society comes from web articles published by various South African news and media outlets. These articles were collected directly from the internet. The sample of articles to collect came from the GDELT project. GDELT is a comprehensive database capturing the world's news on a daily basis (*23*). South African media outlets were determined using two approaches. First, articles published on domains ending with ".za" were considered to be of South African origin. Second, articles published on domains that GDELT's multilingual source-country cross-referencing dataset determined to be of South African origin were considered (*24*). Once identified, those 5,190,260 URLs were scraped, covering a period from 2015 to 2024. Articles which could not be downloaded due to IT policies of the domain were discarded, leaving 1,773,664 full-text media articles from 771 distinct outlets.

Data on the government comes from the Parliamentary Monitoring Group (PMG). The group collects and disseminates materials and notes from more than 50 South African parliamentary committees, among other government documents at the national, not provincial, level (*25*). These documents date between 2000 and 2024. Table 1 details the collected documents.

While paper abstracts and media pieces are of similar length, generally between 200 and 500 words long, the government documents can be much longer. To account for this, they were chunked to the length of the average abstract/media article before being considered in the topic model described in the next section. While a given paper abstract or media article of 300 words may touch on one or two topics, a parliamentary meeting or annual report can run for hundreds of pages, potentially mentioning many dozens of topics. Chunking them ensures a comparable 'document'

unit across the domains. [1]

Henceforth, the term 'document' refers to either an abstract, a media article, or a chunk of a government document. Once the government documents are chunked, the final number of documents used for estimating the topic model is 4,920,490. Chunk size does not affect the results in section 5, as all metrics control for the number of documents in a corpus. Chunking the government documents to the same length as the media articles is only important for the estimation of the topic model described in section 4.1. In that phase, relatively longer government documents could result in relatively fewer topics being identified from them, since one topic is estimated per document.

4 Methodology

Starting with such a large corpus of natural text documents, the question becomes, how can we identify and extract relationships from them on a macro level? For the purposes of this paper, the fundamental piece of information for each document is *what* it is talking about, as opposed to *how* it says it, i.e., elements of diction, syntax, or style. But the challenge is the heterogeneous nature of the domains and the documents that represent their dialog.

For science, there exists a relatively consistent taxonomy of what a given abstract is about in the form of the subject area assigned to the paper . In Scopus' case, each paper is assigned two depths of detail for one or more subject areas. For instance, a paper may be classified as about 'Physics and Astronomy' at a more general scale, and then more specifically as 'Condensed Matter Physics'. Papers' distributions within these subject areas could be analyzed for insights into how scientific discourse is changing, but there are several limitations to that approach. Most obviously, this syntax does not exist for and is not comparable to the other two domains. Additionally, in the case of multiple subject areas assigned to a single paper, there is no way to determine how much the paper is about each of the subject areas. Is it perfectly equally related to physics and chemistry, or mostly about physics and only slightly related to chemistry?

Unsupervised topic modelling addresses both of these issues, providing a consistent and com-

[1] There is the reverse concern, that if a 500 page document reflects, for example, one topic, chunking will create many 'documents' that reflect that topic, potentially skewing results. This concern does not arise because of the way we normalize for the number of documents.

parable measure of what each document is about.

4.1 The topic model

The topic modelling approach used for the analysis was BERTopic, a methodology suitable for unsupervised topic extraction introduced in 2022 (*26*). BERTopic performs three steps: first, the model uses pre-trained language models to embed the documents in a vector space. Word embeddings use artificial neural networks (ANN) to map words or tokens to a high-dimensional vector space. Words and characters can thus be represented by real numbered vectors. This permits one to create distances among them which can to be measured and compared. The vectorization captures semantic relationships between words based on their usage in large text corpora. The coordinates for the words "mountain" and "hill", for instance, will be much closer in the abstracted vector space than the words "mountain" and "computer". This captures the semantic similarity between the words based on the contexts in which they appear. Document and sentence embeddings work similarly, but embed entire sentences and documents to single vectors, rather than individual words (*27*). For our analysis, the *all-MiniLM-L6-v2* embedding model was used. This is a fine-tuning of the MiniLM model introduced in Wang et al. (*28*) and is the most commonly used embedding model, downloaded more than five times as often as the next most-common one on Hugging Face (*29*).

Second, the model clusters the documents based on their spatial positions. To create the clusters, first a sub-step is performed to account for the high dimensionality of the embeddings, which otherwise would complicate and distort distance measures. The Uniform Manifold Approximation and Projection algorithm (UMAP) is used for dimensionality reduction (*30*). Accelerated Hierarchical Density Clustering (HDBSCAN) is then used to cluster the reduced embeddings (*31*).

These clusters of embeddings now represent the identified topics. The final step is to extract topic representations from these document clusters. The model uses a modified, class-based TF-IDF measure (Term Frequency-Inverse Document Frequency) to extract topic-word distributions, which can then be used to understand what topic is being captured in the cluster, and to produce a new embedding for that topic. An example of the top words in the word distribution of a topic is the following: *'climate', 'warming', 'emissions', 'pollution', 'paris', 'carbon', 'change', 'air', '15c',*

'greenhouse'. BERTopic will go no further in terms of assigning a single, semantic title for the topic, that is up to the researcher. In this case, it is clear that the cluster is capturing the topic of climate change.

Due to the size of the dataset in the analysis, one additional step is performed to generate the final topic model. With nearly five million documents, it is infeasible to estimate a single model with all documents without specialized and expensive computing power. To address the issue of scale, many smaller topic models were estimated with batches of 50,000 documents. These topic models were then combined into one using the cosine similarity of the topic vectors (*32*). For instance, say that two small topic models were estimated. When merging them, it is found that they both identified a climate change topic whose vectors are extremely close. The two vectors are then merged into a single climate change topic. However, the second model included some documents on cars, whereas the dataset for the first model did not include any documents on cars. When merging the models, the car topic vector has no vector close to it in the first model's topics, so is retained as a new, separate topic in the topic model after merging.

The merged topic model identified 9,046 topics addressed in the five million documents. The final step was to characterize the documents in terms of the topics. This is done probabilistically. The first step of our process created a vector to describe each document. We compare the cosine similarity of each of those document vectors to each of the topic vectors. The result is a 4,920,490 (documents) x 9,046 (topics) matrix of cosine similarities. For instance, consider a news article talking about a drought in a region. Its cosine similarity to the drought topic is 0.8 (the maximum is 1, indicating the same vector, while the minimum is -1, indicating a perfectly opposite vector), to the climate change topic is 0.6, and to the car topic only 0.1. We can interpret that as a very high probability that the document is talking about drought, a lower, but still high, probability that it is talking about climate change, and a very low chance that it is talking about cars. Having this complete similarity matrix addresses one of the main concerns with using a binary, pre-determined topic classification, such as the Scopus subject areas. In that, we can now determine the degree to which any document is about any of the topics. In our earlier physics chemistry example, we can now determine quantitatively to what degree the abstract is about physics, and to what degree it is about chemistry, or any other topic.

The next section details how this topic model can be used to extract insights and trends and

draw conclusions from the documents.

4.2 Quantifying domain overlap and focus change

With the cosine similarity matrix, we can calculate various time trends and domain characteristics. The first step is scaling the cosine similarities to remove negative values, which complicate the interpretation of our downstream aggregations. The cosine similarities are normalized via equation (1), where $S_C(A, B)$ is equal to the cosine similarity of vectors A and B:

$$S_{norm}(A, B) = \frac{S_C(A, B) + 1}{2} \tag{1}$$

The resultant similarities are now scaled between 0 and 1, with 0 corresponding to opposite vectors, 0.5 corresponding to no relationship between the vectors, and 1 to identical vectors.

With each document assigned a year, we can now observe time trends in topics' popularity. While summing all documents' cosine similarity to a given topic per year can provide some insight, the raw sums are sensitive to the number of documents in a given year. As such, we may observe growth in a topic's similarities not because it is becoming relatively more popular, but simply because more documents were produced in a given year. It also means that comparisons between domains are not possible, since there are more government documents than media documents or science documents.

An alternative metric is the median cosine similarity for a given topic in a given year, which controls for the amount of documents and allows comparisons over time and between domains. An increasing median similarity implies that the corpus is getting closer to the topic over time, i.e., talking about it more.

While median similarity is useful, signals from it may be muted due to the fact that the majority of documents are not about the majority of topics, meaning many topic-year medians are close to 0.5 with small degrees of change; meaningful signals get drowned out. For a more meaningful metric of topics' developments, we should consider the nature of how topics are discussed. In reality, a given document touches on at most a handful of topics and by extension topics related to those. No 300-500 word document can touch on all topics to a meaningful degree. In our matrix of documents and topics (4,920,490 x 9,046) representing the weight of each topic in each document, there are

many entries very close to 0.5. To facilitate interpretation and to make calculations feasible, we trim this matrix by focusing on the top 5% of documents in each topic. That is, for every topic we restrict attention to the documents having the highest affinity to that topic. In effect, in each column of the matrix we set the smallest 4,674,465 entries to zero, leaving as present only those documents that are closest to that topic.

For any given topic, using this top fifth percentile of documents, we can calculate what proportion of them come from each of the three domains. This provides insight to the nature of a topic. Is it a topic that is almost exclusively talked about in government, or is it a well-mixed topic covered in equal degrees by all three domains? Do these proportions change over time? Making comparisons across domains by simple counting will be mis-leading, because in our data the different domains contain different numbers of documents. Thus, we convert the number of documents from domain X^2 appearing in (the top fifth percentile of) a topic to a proportion of the total number of documents in domain X.

Table 2 illustrates this concept, showing what the similarity matrix looks like for a single example topic. The rows represent each of our 4,920,490 documents, sorted by cosine similarity to the topic. The domain column indicates if the document comes from science, media, or government. The cosine similarity column represents each document's similarity to the example topic. Since we are only interested in the top fifth percentile of most similar documents, the cosine similarities of documents after 246,025 (the top fifth percentile of all our documents) are zeroed out; they are not considered for our analysis. We are interested in the mixture of documents that comprise rows 1-246,025 in the table. Hereafter, references to 'document' will refer to these documents in the top fifth percentile.

For instance, say in a given year there are 100 documents from all three domains that constitute the top fifth percentile in terms of similarity for a topic. 50 come from government, 25 from media, and 25 from science. Assume government and media have the same number of documents, while science has half as many. Rather than saying government accounts for 50% of the top documents in terms of similarity (i.e., documents most likely to actually be talking about the topic), media 25% and science 25%, we need to correct for the fact that science has fewer documents overall and so a smaller chance of having documents fall into the top fifth percentile. After correcting for the fact

[2] $X \in$ {science, government, media}

that science has half as many documents as the other two domains, we come up with a proportion of 40% government, 20% media, and 40% science.

We can then combine these three figures into one measure of entropy to calculate how well-mixed or concentrated a topic is in terms of the three domains. From the three metrics, the Shannon entropy score is calculated and normalized to a value between 0 and 100 (*33*). A value of 0 means that the topic is perfectly concentrated; the top fifth percentile is made up exclusively of documents from a single domain. A value of 100 means that the topic is perfectly mixed; each domain constitutes exactly 1/3 of the top fifth percentile of the topic. A single entropy score allows us to easily identify topics that are either well-mixed or concentrated, as well as identify when those relative shares are changing over time, becoming either more mixed or concentrated during certain periods. Important to note is that this metric does not capture commensurate growth in interest across all three domains. I.e., if all three domains increase relative interest at the same rate, entropy will remain the same. It does not capture a topic's universal growth in popularity across all three domains. Rather, it captures relative differences. The logic being that if all three domains increase interest at the same rate, they are reacting simultaneously to the same external developments. Whereas, if a topic starts out dominated by one of the domains, then one or both of the other two domains start growing their relative interest at a faster rate while the dominant domain's interest stagnates, declines, or grows more slowly, it is more likely the dominant domain either originated, amplified, or was already highly aware of the topic.

5 Results

5.1 Characteristics of the domains as a whole

Before looking at topic-specific findings, it is helpful to gain an understanding of the domains on an aggregate level. We first focus on the domains in isolation, independent of the others. A first observation is that they are not homogeneous in terms of how their probability is distributed across the topics. Figure 3 shows the Lorenz curves of the three domains.

The Lorenz curve is a commonly used graphical measure of the degree of concentration of a variable (*34*). In this case, it plots the cumulative proportion of probability of each topic against

the cumulative number of topics. As explained in section 4.2, only the probabilities of the top fifth percentile of each topic were used in calculating the Lorenz curve. Otherwise, the curves would be extremely flat, as most documents are not about most topics. The 45 degree red dotted line represents perfect equality, i.e., every topic has the exact same amount of probability/cosine similarity. A line bowed more towards the bottom right indicates higher concentration, i.e., some topics have a higher proportion of the domain's total probability than others.

The science corpus is the most bowed, while the media and government curves track each other closely. We can interpret that as science being the most 'concentrated' or 'specialized' of the three domains, with fewer topics taking up relatively more of its probability. We can quantify this degree of concentration using the Gini coefficient, where a value of 0 indicates the 45 degree line of the Lorenz curve, i.e., equal probability distribution across all topics. A value of 1 indicates perfect concentration, i.e., all of a domain's probability is concentrated in a single topic. Table 3 displays the Gini coefficient by domain for the period from 2015-2024.

The coefficients bear out the visual observation from the Lorenz curves; science is roughly twice as concentrated as either government or media. In the next section, we will examine individual topics. Though science is the most concentrated of the three domains, its concentration has not been static over time. Figure 4 shows the development of science's Gini coefficient over time.

Science is the only domain which displays a consistent temporal development in one direction. The other two change slightly from year to year, but with no consistent pattern. Between 1995 and 2024, science's Gini coefficient decreased from 0.792 to 0.748, a decrease of about 5%. This suggests that South African science may be slowly becoming less concentrated across a wider range of topics. This is consistent with policy measures and observations about the nature science in South Africa.[3]

We can also examine the topics that have experienced the most growth in terms of relative cosine similarity. These are topics which are receiving relatively increasing attention by a domain, and sheds light on what domains are increasingly focused on at a high level.

For media, the topics receiving increasing attention are dominated by climate and weather-

[3]See for example the 1996 white paper, which discusses the need to strengthen under-developed fields. Or Mouton et al. (*14*) page 41, which shows growing diversity within natural science and a strong growth of output in social science.

related issues. Out of the top 20 fastest growing topics, 7 are related to the issue, covering things like wildfires, water, extreme weather events, natural disasters etc. It is important to note, however, that fastest growing does not mean most popular. For instance, the fastest growing topic for media in terms of probability proportion is a topic dealing with disaster and emergency grants. However, the topic ranks in the bottom quartile in terms of overall probability. That means that, though the topic is growing in popularity, it is still relatively unpopular. The top topics in media in terms of overall probability proportion are dominated by the usual suspects in sports, politics, finance, entertainment and crime. If we take those seven climate-related topics and rank them in terms of total probability proportion, we get an average rank of 4,676 (a rank of 1 would be the most popular topic, 9,046 the least popular). That is, though these topics are among the top 20 fastest growing in media in terms of popularity, they still rank in the bottom half of topics in terms of overall popularity.

For government, there is a clear trend in topics that are garnering increased relative attention. Out of the top 20 fastest growing topics in government, more than half, 11, are related to digitization and technology. Digitization topics are not only growing quickly in government, their absolute level of popularity is also much higher than the fast-growing climate-related topics are in media. If we perform the same calculation, taking those 11 topics' average total probability proportion rankings, we get a value of 2,358. These digitization topics are just outside the top quartile of most popular topics in government.

Science's top growing topics are dominated by climate change and environmental sustainability and adaptation. Of the top 20 fastest growing topics, 13 touch on these topics. Even more than government, these topics are not only fast-growing, but extremely popular as well. Those 13 topics' average total probability proportion ranking is 699, placing them well within the top 10th percentile of most popular topics in science.

We turn now to changes in domains over time relative to each other. Figure 5 shows the total and bilateral aggregate topic entropies of the domains over time.

The entropies are calculated in the following manner: for each year, for every individual topic, we calculate the proportion of the top fifth percentile attributable to each domain, and subsequently the topic-year's entropy scores. From this, we calculate a per-year weighted average entropy score,

where the weights are the relative importance of each topic in the year.[4] The weighted average is used to control for potentially large changes in marginal topics and accurately reflect potentially smaller changes in more popular topics.

In terms of the entropy over all three topics (red curve), we can see a slight upward trend over time. This indicates that, in general, the domains are becoming less specialized across topics over time. That is, across all topics, the documents mentioning them are coming increasingly from two or three of the domains, rather than from a single one. Breaking down the bilateral entropies, we can see what is driving this trend. Between 2015 and 2021, the upward trend was largely driven by an increase in mixing between media and science. After 2021, it was driven mostly by an increase in mixing between government and media.

A second thing to note is the levels of the entropies. Government and media are significantly more mixed than either of the two domains are with science. Science and media remained further apart than science and government until 2021, after which the degree of mixing remained relatively close between science-government and science-media. This result is robust to bootstrapped samples in terms of both topics and observations. I.e., it holds for random subsamples of topics as well as for equally-sized subsamples of the observations from each of the three domains.

5.2 Development of individual topics

General trends are informative, but can mask movements at disaggregated levels: changes in certain topics may be cancelled out by those in others. The necessity of looking more closely into specific topics is demonstrated by Figure 6, which shows all topics sorted by their median annual entropy change between 2015 and 2024.

In this plot, we have all 9,046 topics identified in the topic model on the x-axis, sorted by median annual change in entropy between domains' proportion of the top fifth percentile of that topic; the y-axis shows the average year to year change in entropy for the topic. A lower value indicates a topic that is becoming more concentrated over time, i.e., being talked about increasingly by only one domain. While a higher value indicates a topic that is becoming more mixed over time, i.e., being talked about increasingly by multiple domains. We can see that the vast majority of topics

[4]The relative importance of a topic is that topic's proportion of the sum of all topics' cosine similarity in a given year.

fall between -2 and 2, that is, their entropy generally does not change significantly over time. There are, however, many topics where it does. 675 topics have an annual median entropy change greater than 2 (they are becoming more mixed), while 653 have a value less than -2 (they are becoming more concentrated.

Broadly, there are four dynamics observable in the data, and we will examine examples of all four. First, there are topics that are largely the purview of a single domain, and this relationship is constant over time. Second, there are topics that are mixed, and this relationship is constant over time. Third, there are topics that started out more concentrated, but have become more mixed over time, and fourth, there are topics that experience a temporary acute change in mixing, but where that change decays back to pre-spike levels over time.

Figure 7 shows an instance of the first dynamic, domain concentration.

The plot shows the proportion of documents similar to the topic which come from government, media, or science. The entropy score shows how mixed or concentrated those proportions are. An entropy score of 0 indicates that all the documents in the top fifth percentile (i.e., those most likely to actually be talking about the topic) come from a single domain. An entropy score of 100 indicates that exactly one third of the documents in the top fifth percentile come from each of the three domains. In this plot, which covers the celebrity couple relationships topic, we can see that the top fifth percentile is overwhelmingly made up of media articles, with very few coming from the other two domains. The orange media line near 100% corresponds to the red entropy line below 10. This topic is very concentrated in media, with little change over time. The interpretation is that celebrity couples are mainly talked about in the media and nothing is happening to change this. Other examples of concentrated topics in the media include things like car accidents, music and entertainment, and certain sports topics. Examples of concentrated topics in government include things like financial statements, elements of procedure, budgetary matters, and so on. Examples of concentrated topics in science include specialized academic topics, like lipid metabolism, photophysiology, molecular bonds, etc.

For context, in terms of how many of the topics fall into this paradigm, 21% of the topics get more than 90% of their top fifth percentile (on average between 2015-2024) from a single domain. About half get more than 75%, and 91% get more than 50% of their top fifth percentile from a single domain.

Figure 8 shows an instance of the second dynamic, mixed domain attention.

This topic covers issues of social and political intolerance, containing words like “intolerance”, “hate”, “stigmatising”, “political”, etc. We can see that the proportions remain close to each other over the years, hovering between 30 and 40% for all three domains. The topic’s entropy score is correspondingly close to 100, indicating a high degree of mixing. The interpretation is that intolerance is a topic that is covered in equal measure by the three domains, steady over time. Topics where all three domains are so closely aligned are rare, with no real consistent pattern or overarching characterization. For context, only 6% of topics get at least 20% of their top fifth percentile from all three domains.

More common is alignment between only two of the domains, usually between government and media. 18% of topics get at least 30% of their top fifth percentile from both government and media, including topics like municipal governance and public infrastructure. The figure for science and media is 6%, and for science and government 3%.

Figure 9 shows an instance of the third dynamic, where degree of mixing is not static over time.

The topic deals with mobile applications and the gig economy, including words like “gig”, “apps”, and “phones”. We can see clear a clear time trend in the data, with the topic starting out as the purview of media, but with a steadily increasing share coming from both science and government. As a result, the red entropy line rises significantly over the 10-year period, indicating the topic is becoming more and more mixed. Topics like this can be identified by median year over year entropy change. From Figure 6, we can see that entropy growth from topic to topic changes mostly linearly between -2 and 2, at which point the curve kinks and entropy growth accelerates. If we consider topics with a median entropy growth above 2 as those with a noteworthy degree of change over time, we identify 675, or about 7.5% of all topics.

Looking at these 675 topics, 62% of them start out dominated by media (420). 38% start out dominated by government (254), and only 0.15% start out dominated by science (1). When looking at these 420 topics which start out dominated by media, we see a clear pattern: these topics are dominated by technology, digitization, and data. Of the top 50 (out of the 420) topics that started out dominated by media and have grown fastest in adoption in government and science, four out of five have to do with data and digitization. Topics like AI and robotics in healthcare, cybersecurity, virtual reality, video conferencing technology, telecommunications, and smartphones abound in the

list. If we look at topics that start out dominated by government, we find topics relating to grants, regulation, and government processes (e.g., electoral processes).

This leads to interesting conclusions. Namely, that when topics cross the inter-domain threshold, the common pattern is that they most often originate in media and are usually about technology and digitization. Careful interpretation is required. Crucially, the analysis does not tell us that media originates or initially introduces a topic. Rather, it tells us that a topic is relatively more popular in media, and that subsequently it becomes relatively more popular in government and science. A good example would be artificial intelligence. No one would argue that the media invented artificial intelligence: despite the fact that in the data the AI topic starts as dominated by media, it clearly originated in academia. However, because a topic starts out in academia, that does not mean that it becomes relatively more popular there. It also indicates that a pattern of a topic becoming widely popular in academia, then spilling over to government and media is not particularly common.

For instance, Vaswani et al. (*35*) is commonly cited as the foundational paper for Large Language Models (LLMs). However, when dissecting LLMs' diffusion pattern into all three domains, it was not a matter of the topic being introduced in academia, becoming very popular in academia, then becoming popular in media and government due to its prominence in academia. Rather, they exploded into the public and government consciousness because of their diverse and simple applications for many tasks faced by the general public. This led to a huge amount of attention and discourse around them in the media, which then subsequently increased relative focus in academia and government. I.e., LLMs were initially a niche focus in academia centered around those studying their foundational technology, but after their popularization in the media, they are now frequently mentioned in a wide breadth of academic disciplines. For instance, by economists when analyzing job markets, or education research on learning patterns. This same paradigm can be reasonably applied to many of the domain boundary-crossing topics; initial introduction by academia, magnification and amplification by media if the topic has a broader social relevance, followed by increased relative popularity in academia more broadly and government due to the greater degree of social awareness and interest. We will examine the robustness of these conclusions later in the section.

Finally, Figure 10 shows an instance of the fourth dynamic, where there is a spike in domain overlap that is not sustained.

The topic deals with the mathematical modelling of disease dynamics, containing words like "endemic", "equilibrium", "model", and "disease". From Figure 10 we can see that before 2020, this topic was largely the purview of science, with about 70% of the top fifth percentile coming steadily from academia. However, in 2020, we see a dramatic change. In 2019, media articles accounted for only 15% of the top fifth percentile of the topic. In 2020, that figure jumped all the way up to 44%. Obviously, the increased media attention was due to the COVID-19 pandemic and the sudden relevance of disease modelling to a much broader audience. However, the increased attention was not sustained; disease modelling did not maintain its relative popularity in media. By 2022, the topic had reverted to its pre-pandemic distributions.

Though it is anecdotally self-evident, it is even more apparent from the data that COVID-19 was a singularly unique phenomenon over the time period. No other event was able to coalesce interest across the three domains to such depths (resulting in sudden surges of interest) or breadth (resulting in increased interest across such a wide variety of topics), as evidenced by the vast majority of domain overlap spike topics being COVID-related.

Figure 11 shows another instance of the fourth dynamic, but in a case where the topic was not initially dominated by science. This illustrates an interesting dynamic common to many of the COVID-related topics.

The topic is workplace COVID-19 safety measures, containing terms like 'masks', 'tested', 'employees', etc. We can see that, prior to COVID, this topic was largely dominated by the government. In 2020, however, we see a huge spike in media interest, even overtaking the government's share of the top fifth percentile. We then see, as common to most COVID-related topics, a decay in media interest in subsequent years, by 2023 landing back at its pre-pandemic levels. We did not return to pre-pandemic distributions for government and science, however. We see that as media's share decays, while government does regain some of its lead, science in fact picks up much of the ceded ground. This highlights a key difference in how science and media react to these emergent topics. While media interest spikes very rapidly, but then decays, scientific interest often does not react initially, but rather builds slowly in interest over time in a more durable fashion. The process of writing scientific papers involves an inherent lag. It is difficult to either write a new paper within a year or adjust your in-progress work to respond to an emergent topic. But if the area is ripe for scientific work, we can witness a sustained increase in interest from science long after the media

has moved on.

Another way to understand the impact of COVID is visually, with a network graph, as depicted in Figure 12.

The plots were created in the following manner. A subset of topics was selected to for the sake of readability of the plots. The subset was chosen to include government-focused topics, media-focused topics, science-focused topics, COVID topics, AI-related topics, and water-related topics (salient in the South African context). Once these topics were chosen, the dataset was filtered: we retained a document only if, of the three topics to which it is most similar at least one was in the list above. This produced a more manageable dataset that demonstrates the discourses' dynamics while maintaining legibility.

Nodes of the graph then represent any of a topic, a government document, a media article, or a scientific abstract. Edges represent the cosine similarity between a given topic and a given document. Edges were created between each document and its top three most similar topics. To determine the placement of the nodes, the Scalable Force-Directed Placement (SFDP) algorithm was used (*36*) using all nodes and edges from 2015-2024. While absolute coordinates are not meaningful on the plot, relative positions are. Topics that are closer together share more documents in common, and documents that are closer together share more topics in common.

Topic (red node) coordinates are identical between the two plots in Figure 12. The only difference between them is which documents are plotted. In the case of the top plot, only documents dating from 2015-2019 are plotted, while for the bottom, only those from 2020-2024 are plotted.

In the left-most portion of both plots, we can see two clusters, one dominated by media (orange) and one dominated by government (green). In the case of media, the cluster contains topics and documents about sports. In the case of government, it contains topics and documents about government procedure. In both cases, we see little difference in either density or composition pre- and post-COVID, indicating that these topics were not very affected by COVID. Similarly, we see in the bottom-right corner various specialized academic topics on geology, metallurgy, and materials science. These too experienced little perceptible change after COVID.

Towards the bottom-left, we can see arts and culture, a well-mixed area between media and government. Both pre- and post-COVID, we can see towards the centre a very strong nexus between the three domains centred around the environment. On the right, we can observe the first significant

change between 2015-2019 and 2020-2024. Here, we have topics related to technology and AI. Pre-2020, this area was dominated by media, being predominantly orange in colour. There is a smattering of blue pre-2020, but academic interest has clearly grown substantially since 2020. This is likely not due primarily to COVID, but rather to the increase in popularity of LLMs, a contemporaneous development between 2015-2019 and 2020-2024.

Where we see by far the biggest degree of change, however, is towards the top of the graph, where we find topics related to COVID. Pre-2020, this area of the plot is sparsely populated, though we see a good deal of government dialog about employment as well as media attention on the economy and finance. Scientific literature in this region is not very common pre-2020, save some papers on technical aspects of disease research. Post-2020, however, we see a huge increase in both the density and composition of this region, and the creation of a completely novel nexus involving all three domains. COVID uniquely brought together disparate topics, from the economy to workplace safety to pedagogy to travel to immunology.

5.3 Decomposing entropy changes

An important element to consider when looking at entropy changes is what is driving those changes. If a topic's entropy increases, that can derive from three potential sources:

1. The dominant domain decreases its focus on the topic while the other two domains keep their relative focus, increase it, or also decrease it, but at a slower rate.

2. The dominant domain increases its focus on the topic, but at least one of the other two domains increases it at a faster rate.

3. The dominant domain focus stays the same, but at least one of the other two domains increases it.

To better interpret the entropy dynamics, we can decompose changes in entropy into constituent parts. Given an increase in entropy from year t to $t + 1$ (Δe_{t+1}), we can isolate what proportion of the growth is driven by each of the three domains. If we start with a media-dominated topic, media's growth or decline in relative interest in the topic alone can drive entropy change up or down (ΔMe_{t+1}). If media's relative interest in the topic increases, that will push entropy down (the topic

is already concentrated in media, media is working to further concentrate it). If it decreases, it will push entropy up (media loses interest in the topic, freeing up space in the top fifth percentile for the other two domains, increasing entropy). The inverse is true for the other two domains (ΔGe_{t+1} and ΔSe_{t+1}), since they are starting out as the subordinate domains in the topic. If the other domain's relative interest increases, it will push entropy up, if it decreases, it will push entropy down.

We can combine those three effects to arrive at our total entropy change in formula 2:

$$\Delta Ge_{t+1} + \Delta Me_{t+1} + \Delta Se_{t+1} = \Delta e_{t+1} \tag{2}$$

We can then understand, if a topic is becoming more mixed, how much is due to the subordinate domains increasing their focus (or decreasing it at a slower rate than the dominant domain) versus the the dominant domain decreasing its own (or increasing it at a slower rate than the subordinate domains).

To isolate each of the three impacts on entropy (e.g., ΔGe_{t+1}, government's contribution to overall entropy change), we can use marginal effects. We hold all data constant at time t, and change only the data for a single domain to time $t+1$. For instance, to approximate ΔGe_{t+1}, we hold media and science's data constant at time t, as well as the 95th percentile threshold, but replace government's data with that from $t+1$. We then recalculate the domains' relative shares and entropy to arrive at a new figure for entropy ($\widehat{Ge_{t+1}}$). We subtract that from the initial entropy in time t (e_t) to arrive at $\widehat{\Delta Ge_{t+1}}$, or, the estimated change in entropy moving from t to $t+1$ attributable to only changes in government discourse. We repeat this for the other two domains to obtain the values in formula 3:

$$\widehat{\Delta Ge_{t+1}} + \widehat{\Delta Me_{t+1}} + \widehat{\Delta Se_{t+1}} + \varepsilon = \Delta e_{t+1} \tag{3}$$

Which is similar to formula 2, but with the actual decomposed entropies replaced by their estimates via marginal effects and the addition of the error term ε, to account for the the fact that the estimates sum close to Δe_{t+1}, but not necessarily exactly to Δe_{t+1}. To remove the error term and obtain decomposed entropies that sum exactly to Δe_{t+1}, we can apply formula 4:

$$\Delta e_{t+1}\left(\frac{\sum_{i\in\{G,M,S\}}\widehat{\Delta ie_{t+1}}}{\widehat{\Delta e_{t+1}}}\right)=\Delta e_{t+1} \tag{4}$$

Where $\widehat{\Delta e_{t+1}}$ is equal to $\widehat{\Delta Ge_{t+1}}+\widehat{\Delta Me_{t+1}}+\widehat{\Delta Se_{t+1}}$. The broken out term $\Delta e_{t+1}(\widehat{\Delta Ge_{t+1}}/\widehat{\Delta e_{t+1}})$ etc. is then our scaled, decomposed entropy by domain and what is referred to hereafter as a domain's share of entropy.

The approach can be made clearer by way of example. If we revisit figure 9, the topic of mobile apps and the gig economy, we can apply the methodology to obtain the results in figure 13. The left axis and four lines on the plot are identical to those in figure 9. The new additions are the bar plots, which correspond to the right axis. The stacked bars are the result of applying the entropy decomposition methodology to each year's change in entropy. Breaking it down for two of the years:

- **2016 to 2017:** entropy increases greatly this year, as media's share of the top fifth percentile drops sharply, while the other two domains' shares rise. The large orange bar tells us that a significant proportion of this increase in entropy is driven by changes in media, in this case, a large decrease in relative focus. The smaller blue bar tells us that science has increased its relative focus in the topic, further pushing entropy up. The negative green bar tells us that government, like media, has also decreased its relative focus in the topic, which would, all else equal, reduce entropy. However, the decrease by media and increase by science more than offsets this negative pressure on entropy.

- **2022 to 2023:** entropy decreases this year. The positive blue and green bars tell us that science and media have increased their relative focus on the topic in this year, but the negative orange bar indicates that this increasing pressure is more than negated by negative pressure from media. Media has also increased its relative focus in this year, and at a higher rate that science and government, serving to lower entropy.

We can run the marginal entropy calculation between 2015 to 2024 to get each domain's contribution to the topic's entropy change over the entire period. Doing so, we get a value of 11.3 for government, 13.6 for media, and 8.8 for science. This tells us that an increase in government focus drove 33.5% of the growth in entropy, a decrease in media focus drove 40.4%, and an increase in science focus drove 26.1%.

We can run this calculation for all of the 420 topics that started out as media-dominated and have median annual entropy growth above two. I.e., the media-dominated proportion of the 675 topics at the right end of figure 6. Similarly, we can run it for the 254 topics which started out as government-dominated. Taking the median values for these topics, we arrive at the values in table 4.

The percentages in each column sum to 100%. We can interpret these results in the following manner. When a topic starts out dominated by government, then becomes more mixed over time, the majority of that mixing is driven by media increasingly adopting the topic. Only a tiny percentage is driven by science increasing interest in the topic. By contrast, when a topic starts out dominated by media and then becomes more mixed, it is largely driven by media dropping focus on the topic (i.e., the transient nature of media attention), but is also driven to a significant degree by government and science increasing their focus on those topics.

These findings, coupled with those from figure 5, indicate that we may not indeed have drifted so far from Polanyi's Republic of Science paradigm. Science as a domain among the three is the most separate, contrasting to the tighter linkages between government and media. When domains do come closer together on topics, science is more likely to move towards those initially dominated by media, and much less those dominated by government. This may hint that science does indeed pay attention to topics the general public and wider society care about (assuming presence in media as a proxy), but is resistant to that influence coming from government pressure, supporting Myers' (*13*) findings of a 65% surcharge required to pursue government research. Like Einstein's inspiration for the theory of relativity (a clock tower he saw out the tram window while working in Bern), scientific inspiration seems likelier to come from what scientists observe in the world around them, rather than what is stipulated by government offices (*37*).

6 Conclusion

The COVID-19 pandemic brought acute focus to the government-media-science nexus, highlighting the complexity of each domain's relationship to the others. This coming on the back of years of increasing doubt and change as to the functioning and prioritization of science and its outputs. By analyzing a large text corpus from each of the three domains and quantifying their interactions, we

hope to have added some clarity to the debate around science's behavior and role in broader society.

The results illustrate three key points about the characteristics of and the relationships between government, media, and science. First, government and media are more closely aligned in terms of the topics they cover than either science and government or science and media. This might signify that policy discussions are influenced more by immediate media concerns than by scientific research. For policymakers, this highlights the potential value of mechanisms to integrate scientific evidence into policy debates, especially on complex issues that require long-term perspectives.

Second, relative topic popularity rarely follows a pattern starting with science then moving to media and government. Of the 675 topics with sustained median entropy growth, i.e., that are becoming more cross-domain over time, only 0.15% started out as science-only topics, compared to 62% for media and 38% for government. It is a much more common pattern for a topic to start out with high media attention, and then for government and science to slowly follow. For policymakers, this could be relevant in terms of future research priorities and topic discovery. Media output could be quantitatively analyzed for emerging trends and early identification of what topics could/should become cross-domain.

Finally, it is clear that media is much more reactive and agile, with a shorter attention span, in terms of the topics it covers. Interest in topics can spike and decay quickly, as evidenced by numerous COVID-related topics. Government and especially science are slower to redirect their attention to new topics, but then also slower to drop that attention. This has policy implications for crisis management and long-term strategic planning. During events like the COVID-19 pandemic, the rapid media uptake of scientific topics like disease modelling was critical for public awareness, but its subsequent decline highlights a challenge for sustained public engagement. Policymakers should be aware of the transient nature of media attention and develop robust, enduring channels for scientific advice and public education that can outlast immediate crises.

It also highlights how media is inherently reactive; when topics gain popularity there, they are already more broadly socially relevant. As a result, any attention-led policymaking approach will be chronically behind and ill-equipped to handle novel developments. Further, the results show science to be particularly resistant to following government's lead. Among government-dominated topics that become more mixed over time, only a small portion of this increasing mixture comes from science increasing its focus to the government-favored topics, the majority comes from media. This

is a case for a continued emphasis on fundamental research, rather than micro agenda setting. Many of the topics that diffused from media then to science and government nevertheless had their start as scientific research. Natural language processing was a niche scientific discipline until it suddenly wasn't. The media did not identify a social need for LLMs, talk widely about the topic, catch the attention of government, who then stoked scientific progress in the field and the subsequent desired social impact by putting conditions on their research grants. Rather, researchers performed fundamental research on the topic and at some point obtained an output that had wide-reaching social impact, stoking a high degree of media attention that was later reflected back in government and other aspects of science, who sought to further develop, utilize, and regulate the technology.

Though these results are already interesting, there is large scope for further research using these methods and data. Expanding the analysis to additional geographies is one important option. How many of these results and dynamics are specific to the South African context, and how many are applicable to different countries? It would also be useful to disaggregate the scientific papers by research discipline and see how the dynamics differ as a result. Which scientific disciplines are the most reactive or static? Which have the highest overlap with the other domains? The topics could also be aggregated via hierarchical clustering to get insights to dynamics around how the domains talk about different aspects of topics. For instance, at the individual topic level there might not be much overlap between media and science for the network spectrum topic and the smartphone topic, but when aggregating the two to a digital connectivity topic, we see new relationships not visible at the disaggregated level.

In an increasingly complex and interconnected world facing new collective challenges every year, it is crucial we understand quantitatively how different aspects of society operate and interact.

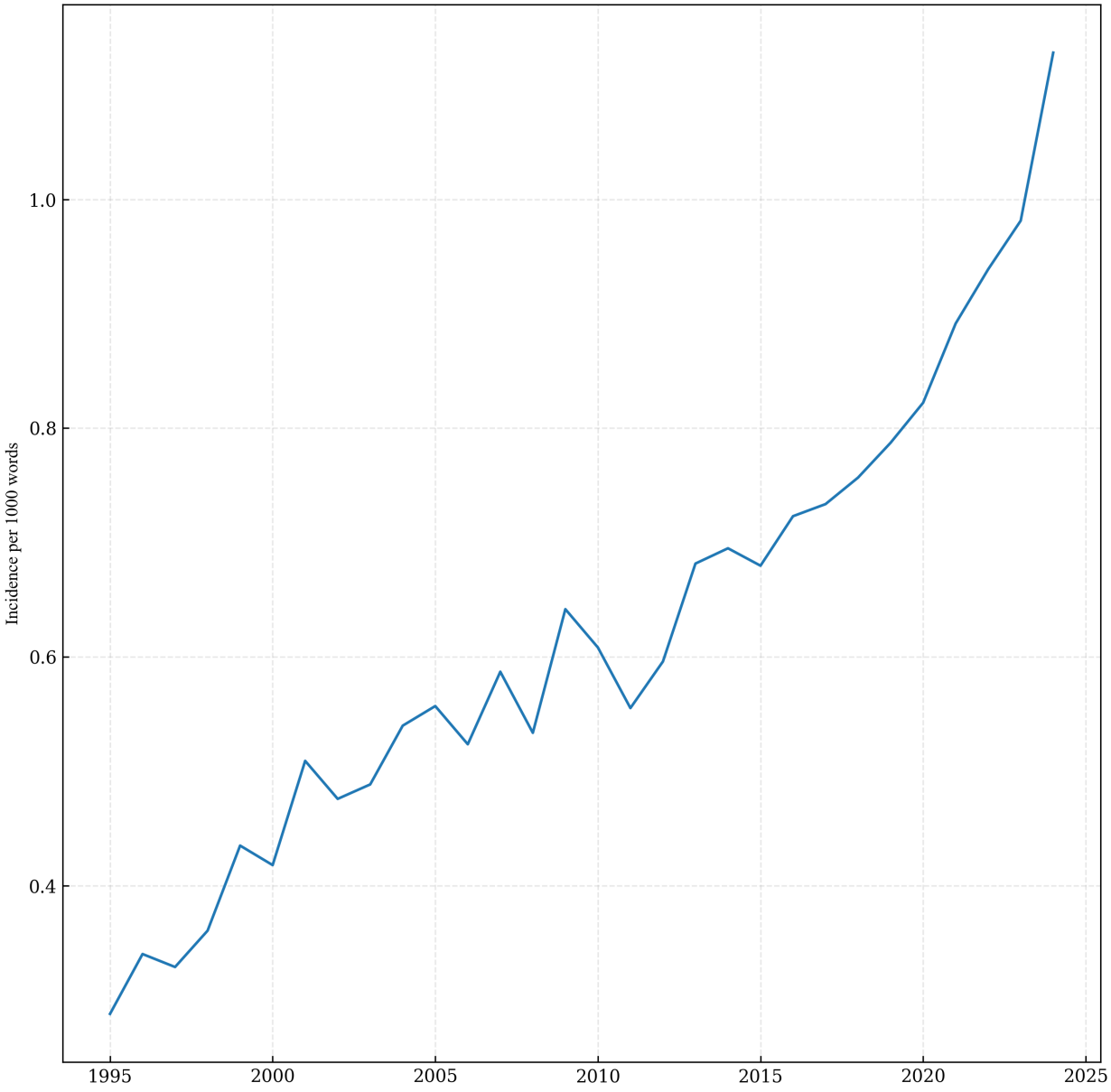


Figure 1: **Incidence of the terms 'impact' or 'impacts' per 1000 words in South African scientific abstracts, 1995-2024.**

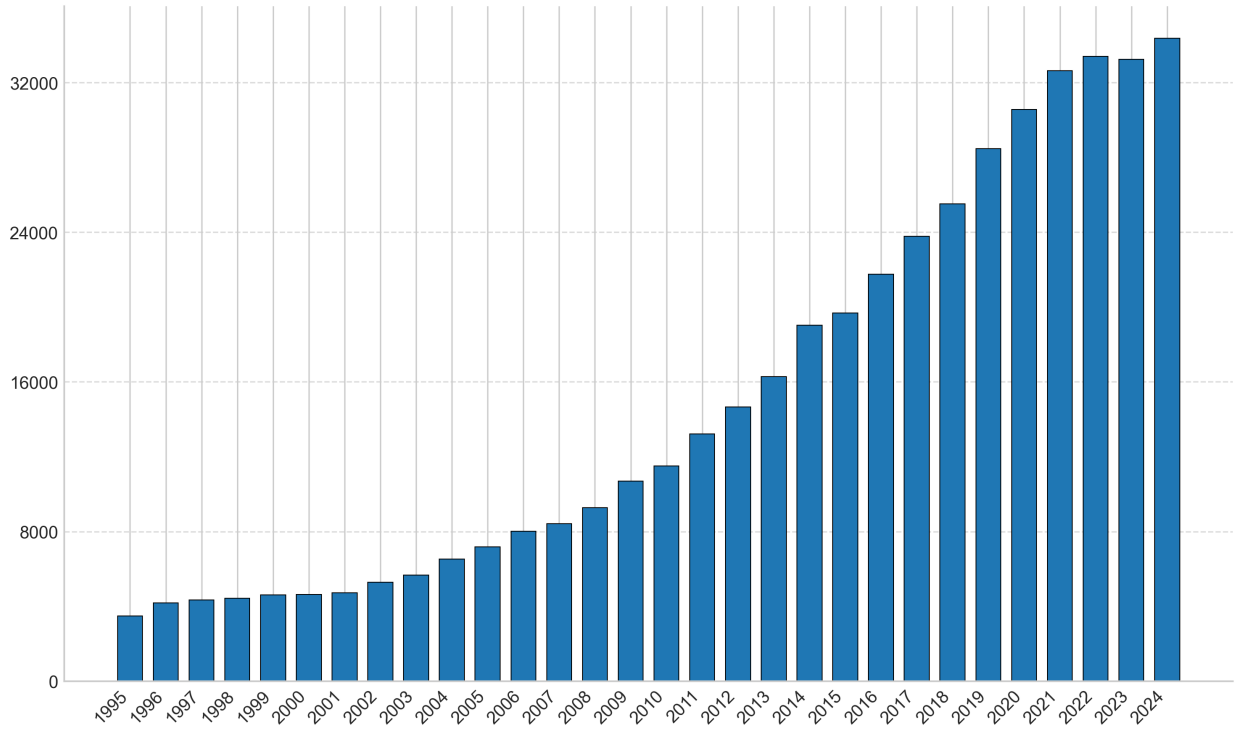


Figure 2: **Number of South African institution-affiliated abstracts by year.**

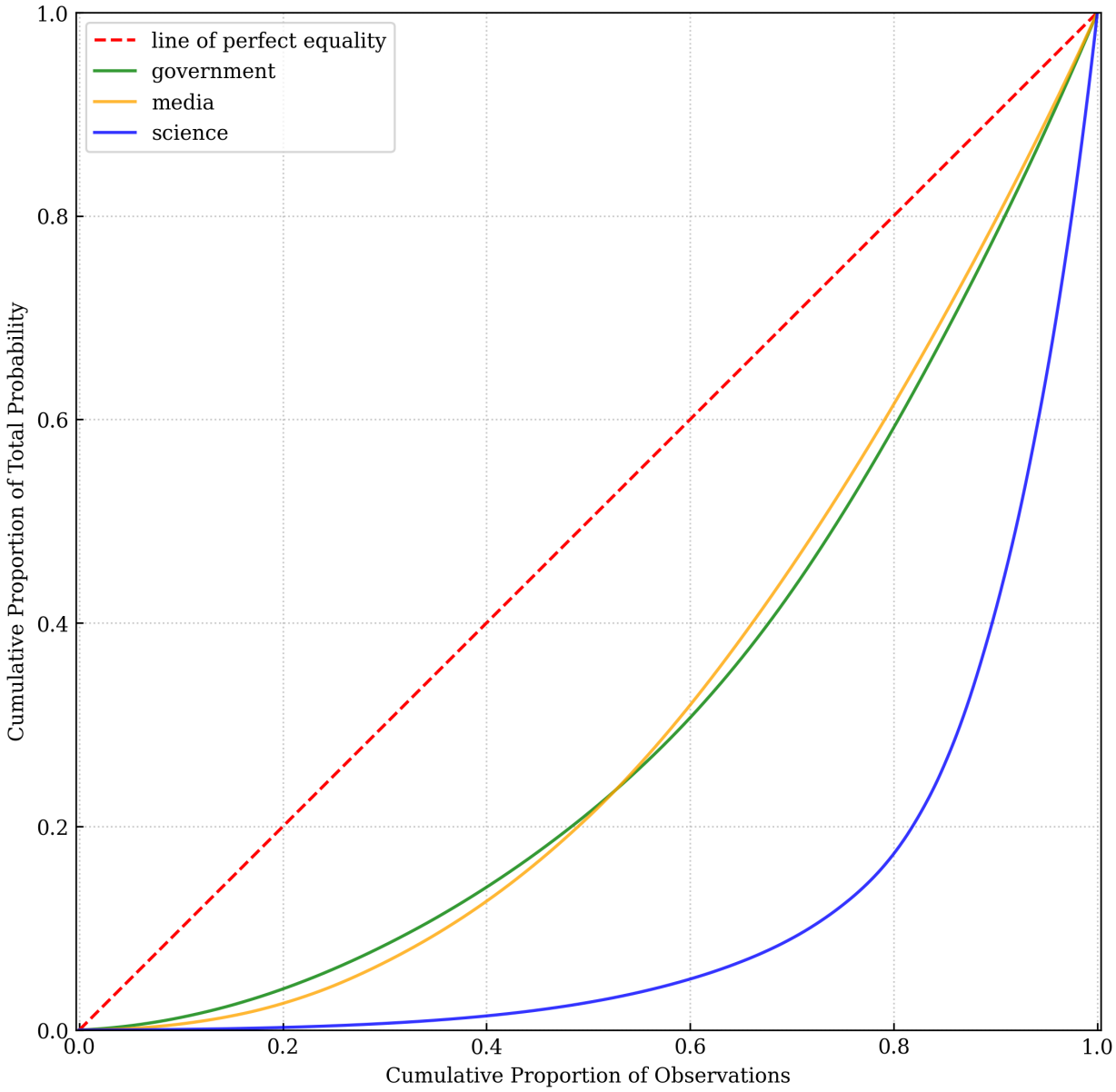


Figure 3: **Topic distribution Lorenz curves of the three domains, 2015-2024, top fifth percentile.** The red 45 degree line represents total equality, i.e., equal probability for all topics. A line bowed towards the bottom right indicates higher probability concentration in some topics.

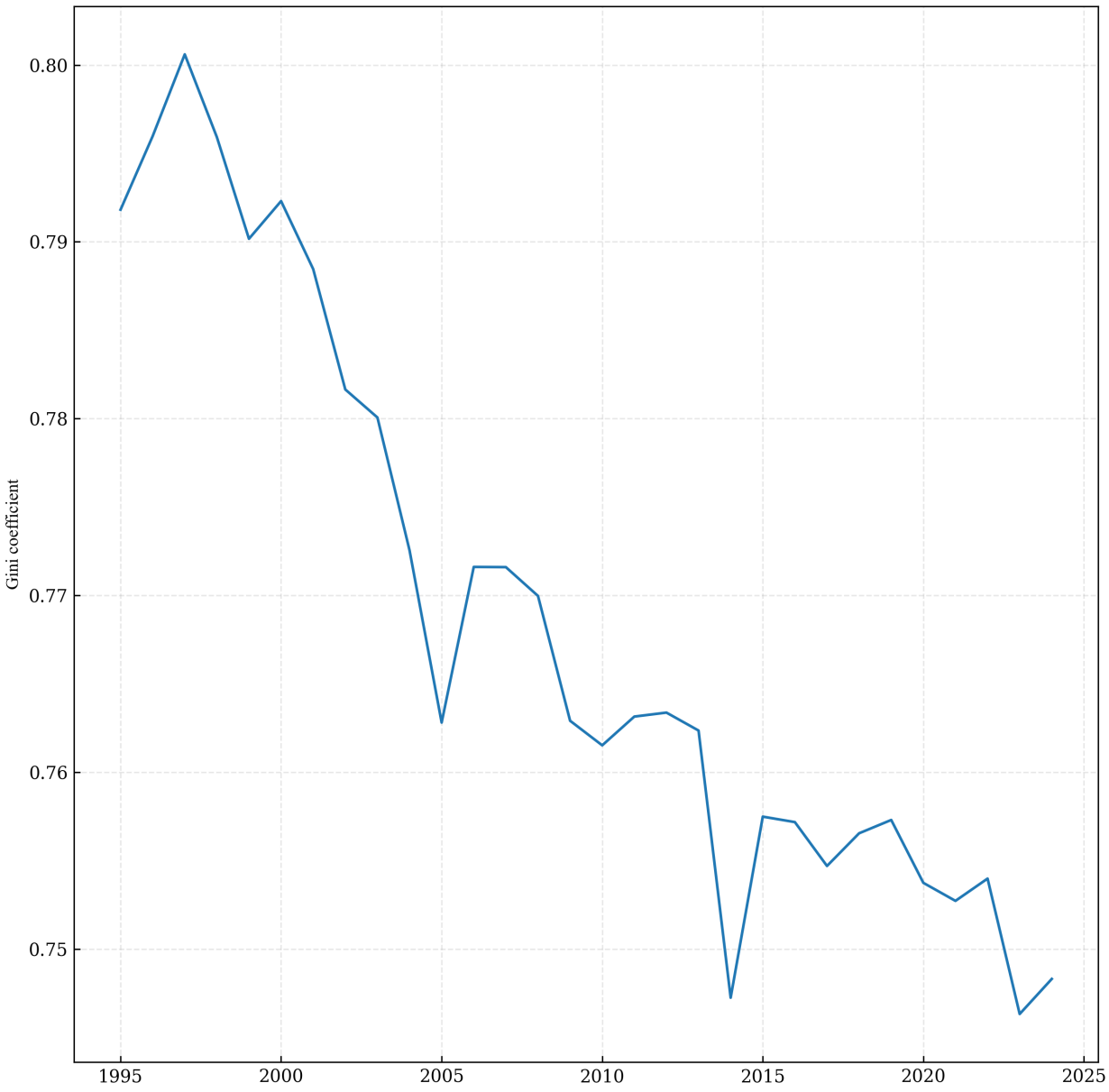


Figure 4: **Gini coefficient of scientific abstracts, 1995-2024.**

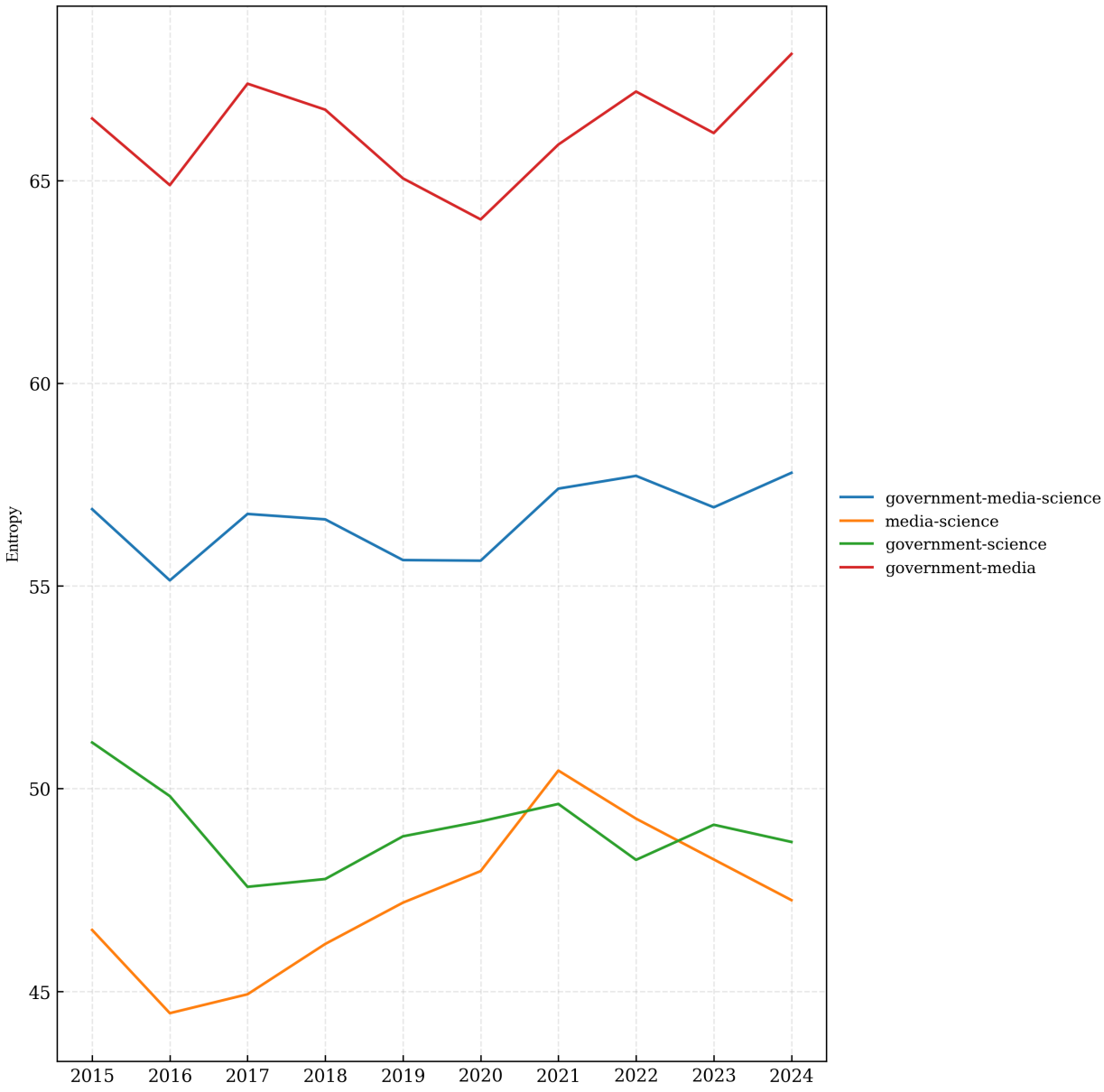


Figure 5: **Entropy between domains' proportion of the top fifth percentile of all topics, 2015-2024.** A value of 0 indicates perfect concentration; the top fifth percentile of all topics is perfectly concentrated in a single domain. A value of 100 indicates perfect mixing; the top fifth percentile of all topics is equally distributed across domains.

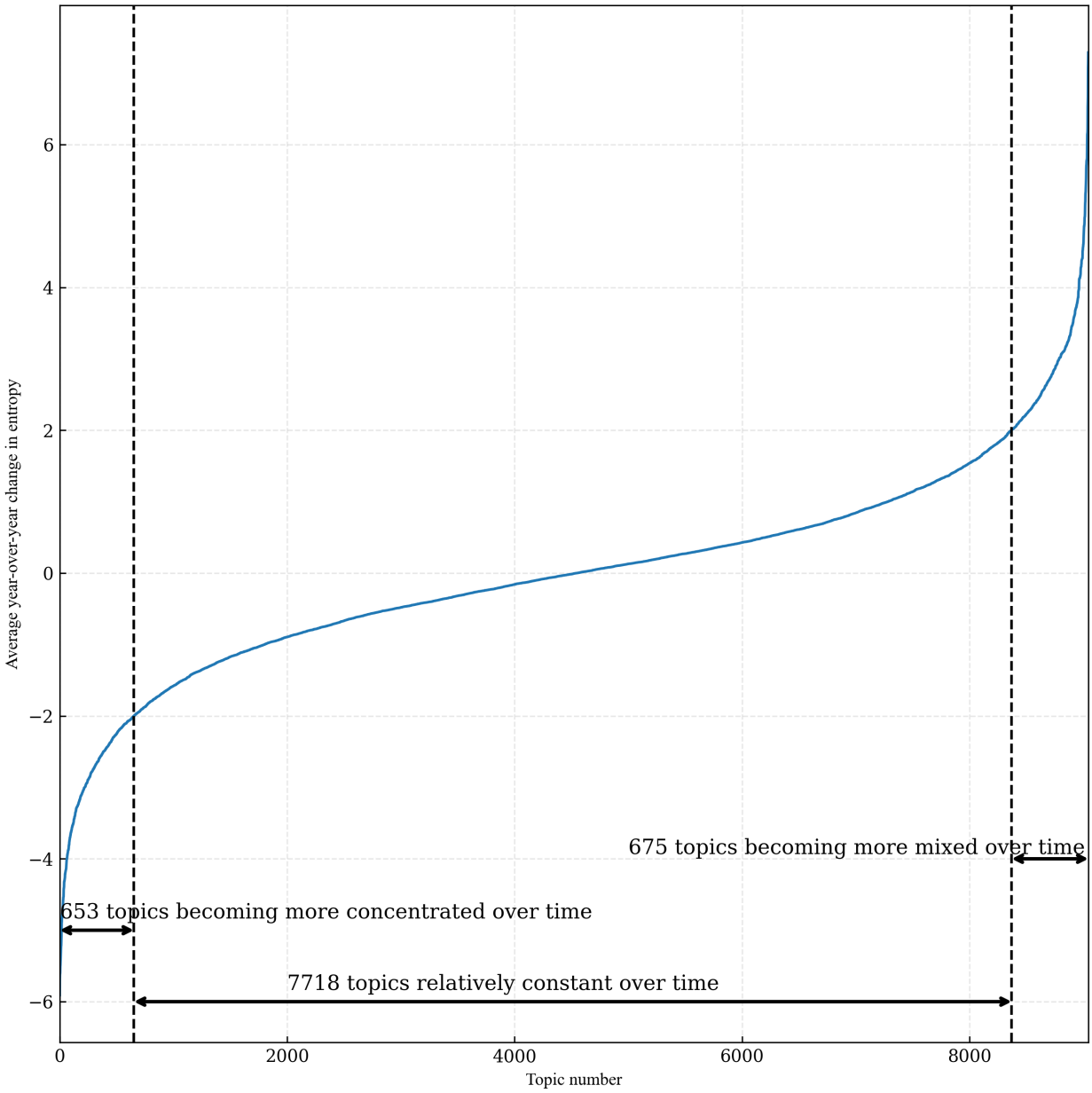


Figure 6: **Median annual change in entropy between domains' proportion of the top fifth percentile of all topics, 2015-2024.** Each point on the x-axis represents a topic, while the corresponding y-axis value shows the median annual change in entropy of that topic. A lower value indicates a topic that is becoming more concentrated (being talked about by only one domain) over time, while a higher value indicates a topic that is becoming more mixed (being talked about by multiple domains) over time.

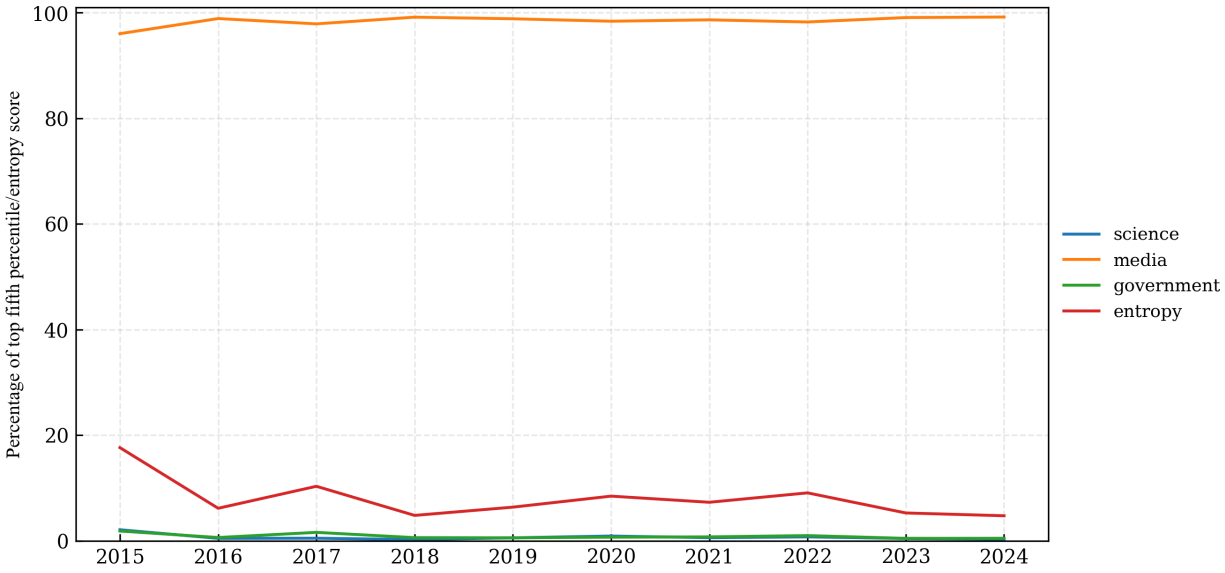


Figure 7: **Topic: celebrity couple relationships.** Of the top fifth percentile of documents similar to this topic, proportion thereof coming from government, media, or science, and entropy of those proportions. An entropy value of 0 means that the top fifth percentile of documents similar to the topic come exclusively from one domain, while an entropy value of 100 means that exactly one third of them come from each domain.

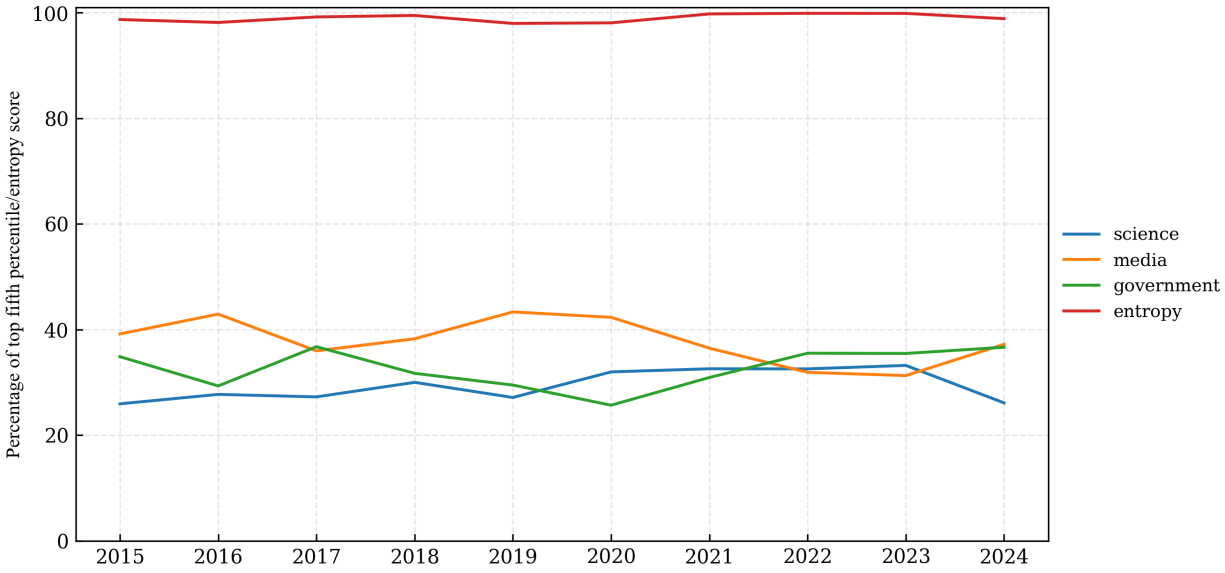


Figure 8: **Topic: social and political intolerance.** Of the top fifth percentile of documents similar to this topic, proportion thereof coming from government, media, or science, and entropy of those proportions. An entropy value of 0 means that the top fifth percentile of documents similar to the topic come exclusively from one domain, while an entropy value of 100 means that exactly one third of them come from each domain.

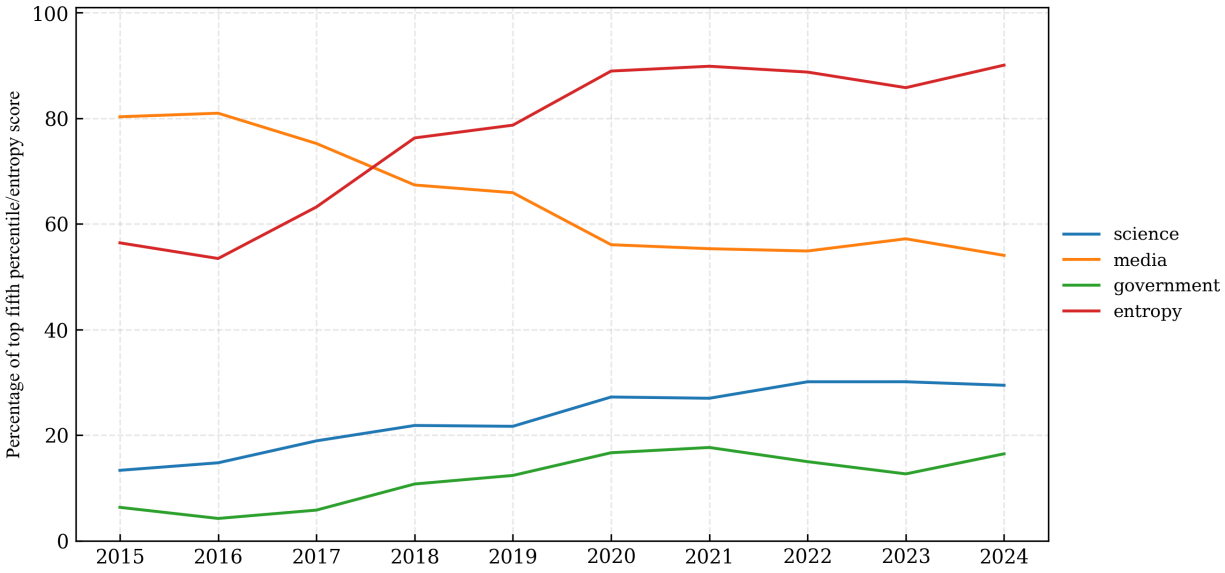


Figure 9: **Topic: mobile apps and the gig economy.** Of the top fifth percentile of documents similar to this topic, proportion thereof coming from government, media, or science, and entropy of those proportions. An entropy value of 0 means that the top fifth percentile of documents similar to the topic come exclusively from one domain, while an entropy value of 100 means that exactly one third of them come from each domain.

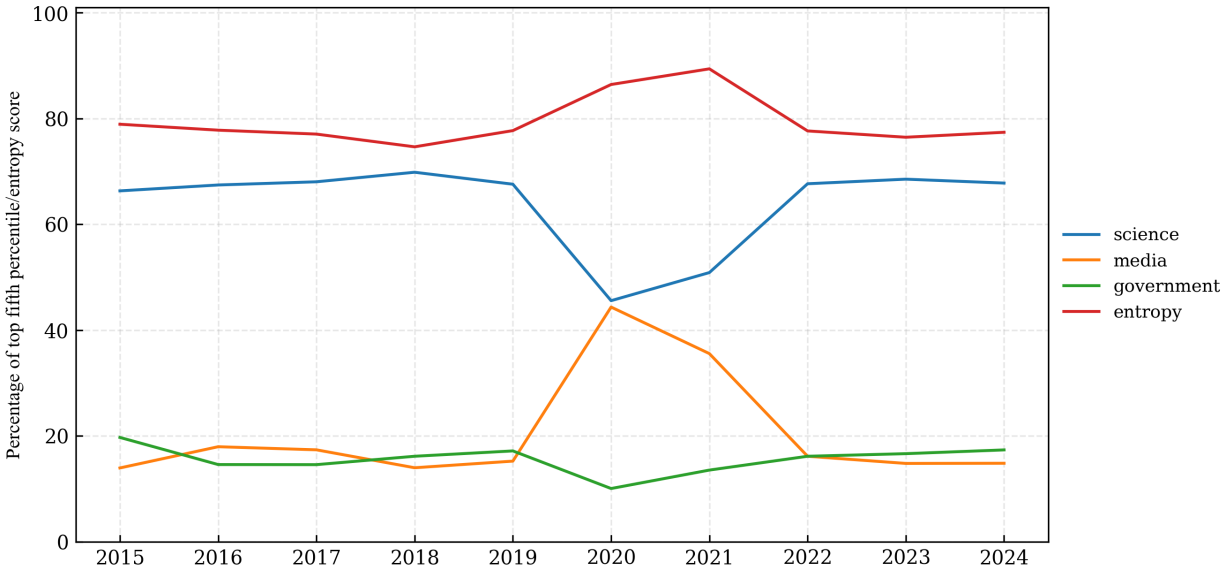


Figure 10: **Topic: mathematical modelling of disease dynamics.** Of the top fifth percentile of documents similar to this topic, proportion thereof coming from government, media, or science, and entropy of those proportions. An entropy value of 0 means that the top fifth percentile of documents similar to the topic come exclusively from one domain, while an entropy value of 100 means that exactly one third of them come from each domain.

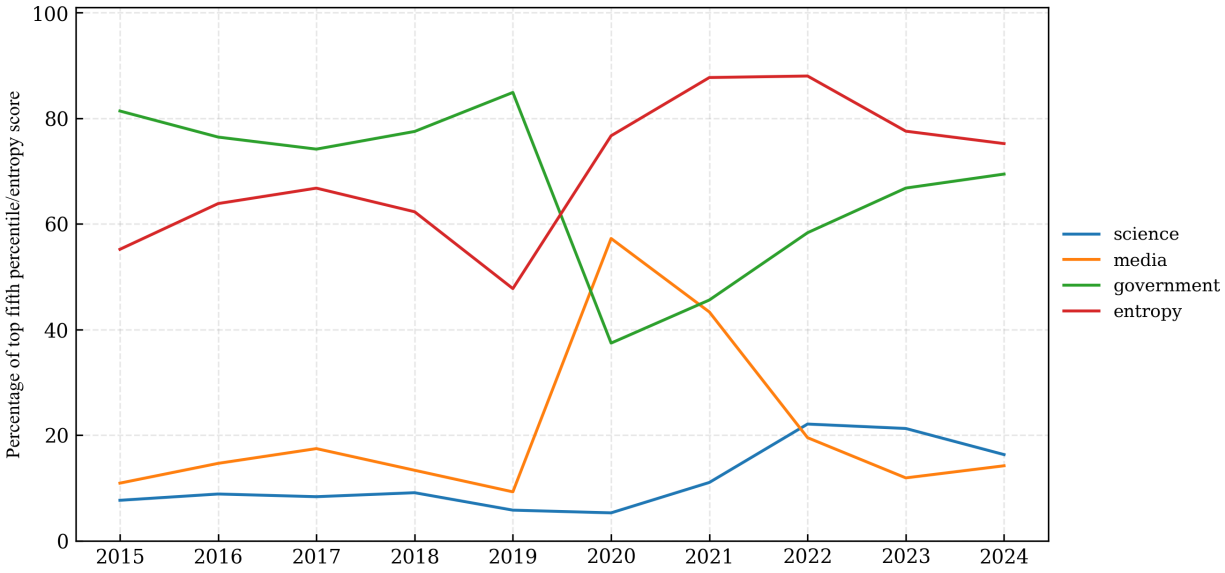


Figure 11: **Topic: workplace COVID-19 safety measures.** Of the top fifth percentile of documents similar to this topic, proportion thereof coming from government, media, or science, and entropy of those proportions. An entropy value of 0 means that the top fifth percentile of documents similar to the topic come exclusively from one domain, while an entropy value of 100 means that exactly one third of them come from each domain.

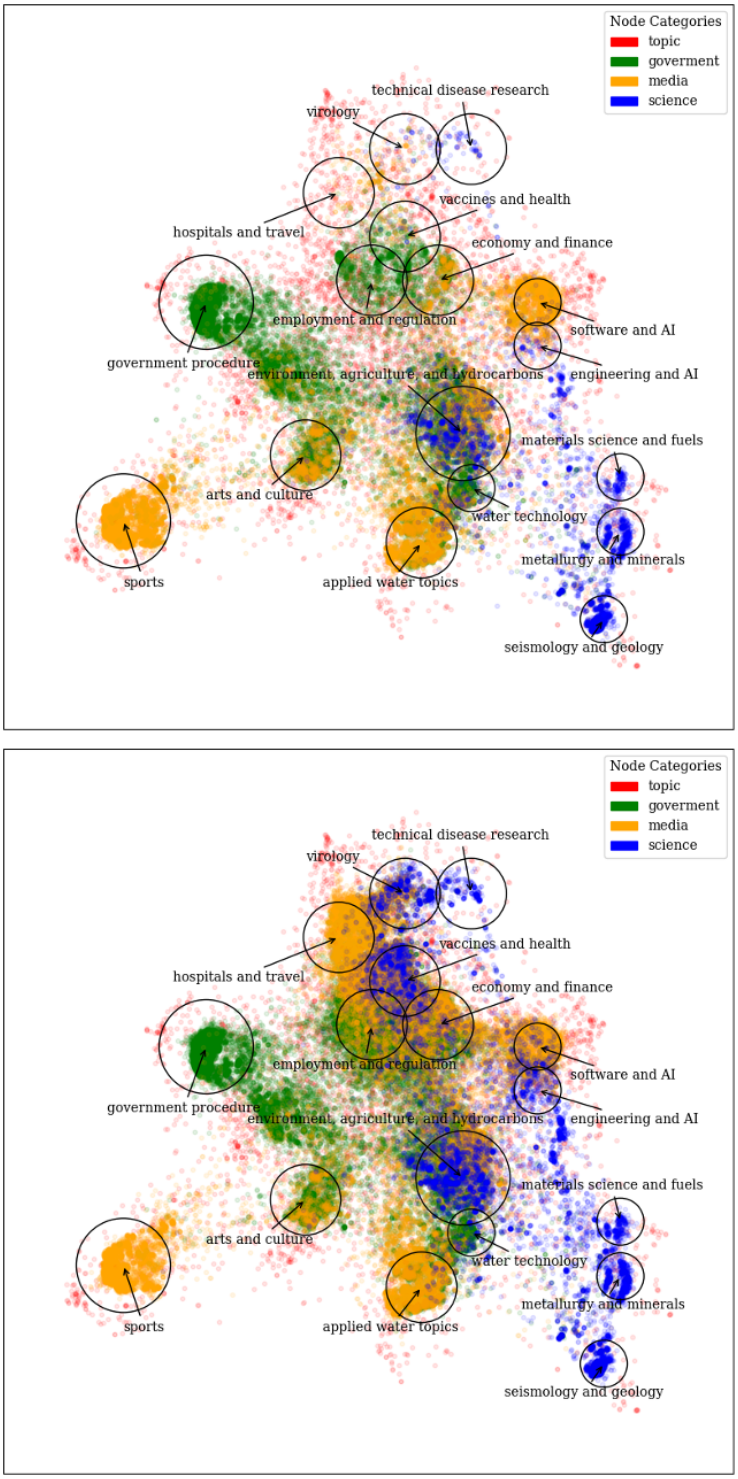


Figure 12: **Network graph of topics and documents, topic subset 2015-2019 (top) and 2020-2024 (bottom).** Nodes represent topics or document, edges (hidden for readability) represent cosine similarity between a topic and document.

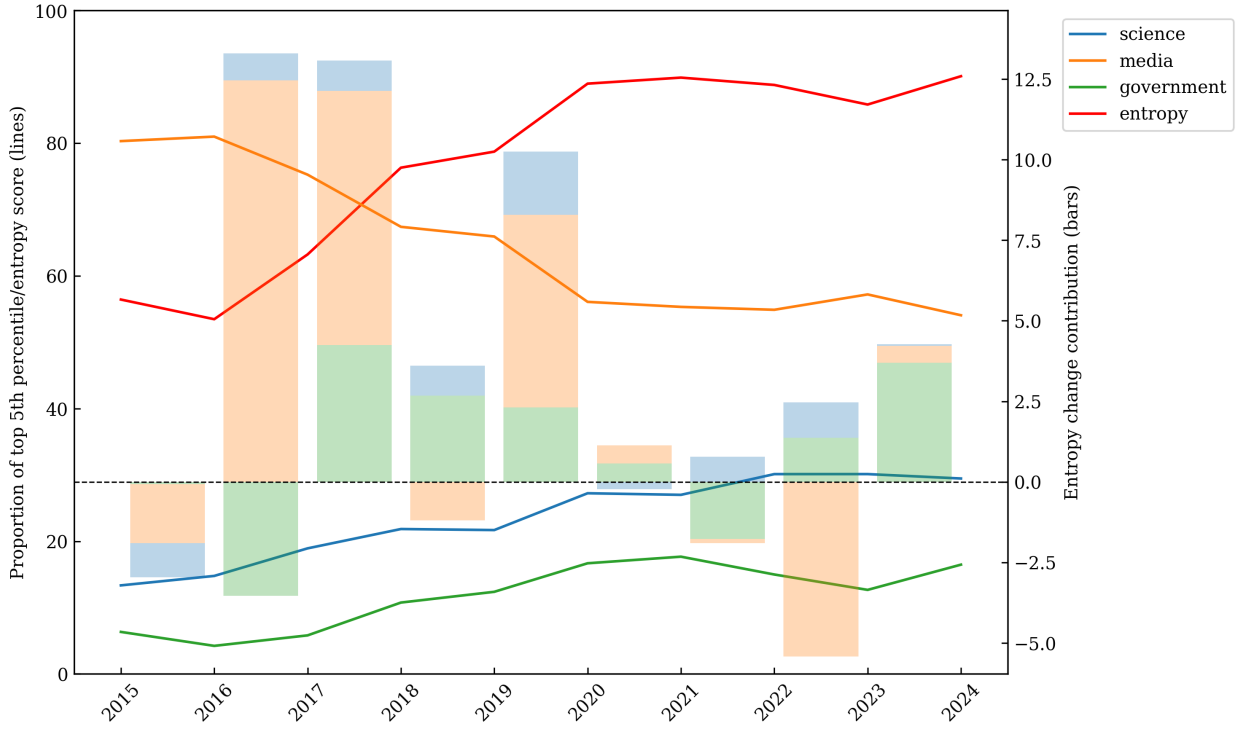


Figure 13: **Topic: mobile apps and the gig economy.** Left axis: of the top fifth percentile of documents similar to this topic, proportion thereof coming from government, media, or science, and entropy of those proportions. An entropy value of 0 means that the top fifth percentile of documents similar to the topic come exclusively from one domain, while an entropy value of 100 means that exactly one third of them come from each domain. Right axis: contributions of the three domains to changes in entropy.

Table 1: **Collected South African government documents.**

Description	Time period	Number of documents
Annual performance plans from government departments	2018-2024	1,754
Annual reports from government departments	2015-2023	18,171
Bills introduced to parliament	2000-2024	703
Calls for comment to the general public	2013-2024	1,467
Minutes for parliamentary committee meetings	2007-2024	20,054
Parliamentary hansards	2000-2024	1,953
Parliamentary media briefings	2008-2024	639
Policy documents laid before parliament	2015-2024	1,247
Questions posed to parliament	2022-2024	5,698

Table 2: Cosine Similarity by Domain to Topic *X*

Document n	Domain	Cosine similarity between document and topic
1	science	0.785
2	government	0.784
3	media	0.776
4	media	0.757
5	government	0.676
...		
246,024	government	0.656
246,025	media	0.655
246,026	science	0
246,027	media	0
...		
4,920,489	science	0
4,920,490	government	0

Table 3: **Gini coefficient by domain, 2015-2024.**

Domain	Gini coefficient
government	0.384
media	0.380
science	0.760

Table 4: **Median entropy contribution by domain to topics with growing entropy**

Domain	Government-dominated	Media-dominated
government	3.9 (30.5%)	5.7 (34.5%)
media	8.6 (67.2%)	7.2 (43.7%)
science	0.3 (2.3%)	3.6 (21.8%)

References and Notes

1. NSF, The National Science Foundation: A Brief History (1994), https://www.nsf.gov/about/history/nsf50/nsf8816.jsp.

2. J. R. Steelman, *Science and public policy*, vol. 1 (US Government Printing Office) (1947).

3. M. Polanyi, J. Ziman, S. Fuller, The Republic Of Science: Its Political And Economic Theory Minerva, I(1) (1962), 54-73. *Minerva* **38** (1), 1–32 (2000), publisher: Springer, http://www.jstor.org/stable/41821153.

4. R. Mandt, K. Seetharam, C. Chang, Federal R&D funding: the bedrock of national innovation. *MIT Science Policy Review* **1**, 44–54 (2020), doi:10.38105/spr.n463z4t1u8.

5. E. Hazelkorn, Public trust and accountability: a time of special challenge (2018), https://www.researchcghe.org/blogs/2018-02-07-public-trust-and-accountability-a-time-of-special-challenge/.

6. M. Brenan, Americans' Confidence in Higher Education Down Sharply (2023), https://news.gallup.com/poll/508352/americans-confidence-higher-education-down-sharply.aspx.

7. W. Broad, Trump Seeks to Cut Basic Scientific Research by Roughly One-Third, Report Shows (2025), https://www.nytimes.com/2025/07/10/science/trump-science-budget-cuts.html.

8. W. H. Lambright, Government and Science: A Troubled, Critical Relationship and What Can Be Done about It. *Public Administration Review* **68** (1), 5–18 (2008), publisher: [American Society for Public Administration, Wiley], http://www.jstor.org/stable/25145570.

9. L. Godfrey, N. Funke, C. Mbizvo, Bridging the science-policy interface : a new era for South African research and the role of knowledge brokering : research article. *South African Journal of Science* **106** (5), 1–8 (2010), _eprint: https://journals.co.za/doi/pdf/10.10520/EJC97040, doi:10.10520/EJC97040, https://journals.co.za/doi/abs/10.10520/EJC97040.

10. L. Bornmann, Measuring the societal impact of research. *EMBO reports* **13** (8), 673–676 (2012), _eprint: https://www.embopress.org/doi/pdf/10.1038/embor.2012.99, doi:https://doi.org/10.1038/embor.2012.99, https://www.embopress.org/doi/abs/10.1038/embor.2012.99.

11. J. B. Holbrook, S. Hrotic, Blue skies, impacts, and peer review. *Roars Trans.* **1** (2013), https://api.semanticscholar.org/CorpusID:29613083.

12. B. Bozeman, C. Boardman, Broad Impacts and Narrow Perspectives: Passing the Buck on Science and Social Impacts. *Social Epistemology* **23** (3-4), 183–198 (2009), publisher: Routledge, doi:10.1080/02691720903364019, https://doi.org/10.1080/02691720903364019.

13. K. Myers, The Elasticity of Science. *American Economic Journal: Applied Economics* **12** (4), 103–34 (2020), doi:10.1257/app.20180518, https://www.aeaweb.org/articles?id=10.1257/app.20180518.

14. J. Mouton, *et al.*, *The state of the South African research enterprise* (2019).

15. J. Priem, H. A. Piwowar, B. M. Hemminger, Altmetrics in the wild: Using social media to explore scholarly impact (2012), https://arxiv.org/abs/1203.4745, _eprint: 1203.4745.

16. L. Bornmann, Scientific Revolution in Scientometrics: The Broadening of Impact from Citation to Societal, pp. 347–359 (2016), doi:10.1515/9783110308464-020.

17. D. S. G. Team, Data Study Group Final Report: Defence Science & Security Laboratory (DSTL) Using Topic Modelling to Discover New Trends in the Scientific Literature (2023), doi:10.5281/zenodo.10033237, https://doi.org/10.5281/zenodo.10033237.

18. V. Tshitoyan, *et al.*, Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571** (7763), 95–98 (2019), doi:10.1038/s41586-019-1335-8, https://doi.org/10.1038/s41586-019-1335-8.

19. L. Dias, M. Gerlach, J. Scharloth, E. Altmann, Using text analysis to quantify the similarity and evolution of scientific disciplines. *Royal Society Open Science* **5** (2018), doi:10.1098/rsos.171545.

20. L. Leydesdorff, The Triple Helix - -University-Industry-Government Relations: A Laboratory for Knowledge Based Economic Development. *Glycoconjugate Journal - GLYCOCONJUGATE J* **14**, 14–9 (1995).

21. V. H. Smith, *Measuring the benefits of social science research*, Impact assessments 2, International Food Policy Research Institute (IFPRI) (1998), https://ideas.repec.org/p/fpr/impass/2.html.

22. L. Bornmann, R. Haunschild, K. Boyack, W. Marx, J. C. Minx, How relevant is climate change research for climate change policy? An empirical analysis based on Overton data. *PLOS ONE* **17** (9), e0274693 (2022), publisher: Public Library of Science (PLoS), doi:10.1371/journal.pone.0274693, http://dx.doi.org/10.1371/journal.pone.0274693.

23. T. G. Project, The GDELT Story (2022), https://www.gdeltproject.org/about.html.

24. T. G. Project, Multilingual Source-Country Crossreferencing Dataset (2015), https://blog.gdeltproject.org/multilingual-source-country-crossreferencing-dataset/.

25. P. M. Group, What is the Parliamentary Monitoring Group? (2024), https://pmg.org.za/page/what-is-pmg.

26. M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure (2022), https://arxiv.org/abs/2203.05794, _eprint: 2203.05794.

27. Q. V. Le, T. Mikolov, Distributed Representations of Sentences and Documents (2014), https://arxiv.org/abs/1405.4053, _eprint: 1405.4053.

28. W. Wang, *et al.*, MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers (2020), https://arxiv.org/abs/2002.10957, _eprint: 2002.10957.

29. HuggingFace, all-MiniLM-L6-v2 (2025), https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

30. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (2020), https://arxiv.org/abs/1802.03426, _eprint: 1802.03426.

31. L. McInnes, J. Healy, Accelerated Hierarchical Density Based Clustering, in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (IEEE) (2017), pp. 33–42, doi:10.1109/icdmw.2017.12, http://dx.doi.org/10.1109/ICDMW.2017.12.

32. M. Grootendorst, merge_models (2025), https://maartengr.github.io/BERTopic/api/bertopic.html#bertopic._bertopic.BERTopic.merge_models.

33. C. E. Shannon, A mathematical theory of communication. *The Bell System Technical Journal* **27** (3), 379–423 (1948), doi:10.1002/j.1538-7305.1948.tb01338.x.

34. N. C. Kakwani, Applications of Lorenz Curves in Economic Analysis. *Econometrica* **45** (3), 719–727 (1977), publisher: [Wiley, Econometric Society], http://www.jstor.org/stable/1911684.

35. A. Vaswani, *et al.*, Attention Is All You Need (2017), https://arxiv.org/abs/1706.03762, _eprint: 1706.03762.

36. Y. Hu, Efficient and High Quality Force-Directed Graph Drawing. *Mathematica Journal* **10**, 37–71 (2005).

37. P. Galison, *Einstein's Clocks, Poincaré's Maps: Empires of Time* (W.W. Norton, New York) (2003).