Research Article

An Information Geometry Approach to Analyzing Topic Evolution in Scientific Networks: From Physics to International Relations

Artem Chumachenko[1]

1. University of Warsaw, Poland

This study presents a novel methodology for analyzing the evolution of scientific topics through the geometric framework of information spaces. Using mutual entropy-based distance metrics, the approach reveals dynamic relationships between scientific concepts over time, surpassing the capabilities of traditional keyword-based analyses. The framework quantifies the creative influence of publications linked to knowledge brokers by measuring the relative compression these agents induce on the geometry of knowledge networks. Applied to topics derived from ArXiv and JSTOR datasets, the methodology identifies patterns of topic evolution and evaluates the impact of key agents, such as publishers, journals, and countries. The findings offer actionable insights for strategic planning by academic journals, funding agencies, and research institutions, facilitating data-driven decision making to promote emerging research trends and interdisciplinary collaborations.

Corresponding author: Artem Chumachenko, a.chumachenko@uw.edu.pl

1. Introduction

The field of scientific research is in a constant state of flux. Keeping up to date on emerging trends and hot topics is essential for journals, researchers, and policy makers to navigate the rapidly evolving landscape of knowledge production[1][2]. The exponential growth of scientific literature and the increasingly interdisciplinary nature of research have introduced unprecedented complexity to scientific discourse. Although traditional methods such as bibliometric analysis have been

instrumental in tracking trends, they often do not capture the intricate, dynamic interactions between concepts, disciplines, and research communities[3][4]. This complexity requires the development of innovative methodologies that can complement and extend the capabilities of existing approaches.

Recent advances in information theory offer a promising pathway to understand and quantify the underlying structures of scientific knowledge[5][6]. Using computational topology and information-theoretic metrics, researchers can explore the geometry and topology of knowledge networks. These approaches, which leverage methods such as persistent homology and simplicial complex construction, have been effectively applied to analyze co-occurrence networks, high-dimensional feature spaces, and text corpora, demonstrating their ability to uncover hidden relationships, detect structural deficiencies, and bridge knowledge gaps[7][8][9][10]. The study of information topology enhances our understanding of knowledge production and provides actionable insights into the diffusion and evolution of information across disciplines.

This study introduces a novel methodology to quantify the influence of knowledge brokers on the dynamics of scientific topics represented as knowledge networks. Specifically, we use the variation of information (VI) distance metric[11], an information-theoretic measure, to assess multiscale geometric topological properties and higher-order interactions within knowledge dynamics. Unlike non-Euclidean metrics, such as those derived from Kullback-Leibler or more general Bregman divergences, which require customized tools as described in[12], the Euclidean nature of the VI metric enables the use of well-established computational frameworks, simplifying its integration into existing pipelines. Our approach offers a straightforward interpretation of topic dynamics as simplex volume compression, facilitating the identification and analysis of knowledge brokers whose influence bridges gaps within knowledge networks.

To the best of the author's knowledge, there are currently no established methods that enable a direct comparison with the proposed framework, particularly in terms of quantifying the impact of knowledge brokers, such as journals, institutions, or authors, on the knowledge network constructed from scientific concepts. Traditional bibliometric approaches, such as citation networks, co-citation analysis, and bibliographic coupling, focus primarily on document-level interactions rather than the evolution of semantic structures within concept-based networks. Although some semantic network and topic modeling methods capture thematic shifts and conceptual evolution, they generally do not attribute changes in knowledge structures to specific agents or quantify their relative contributions. This gap underscores the novelty of the proposed framework, which integrates velocity matrices and

spanning-tree analysis to assess how agents influence the dynamics of conceptual linkages, filling a crucial void in existing scientometric methodologies.

By applying this methodology to datasets from ArXiv and JSTOR, our objective is to demonstrate its utility in quantifying the effects of knowledge production, assessing the influences of different knowledge brokers such as countries and publishers by identifying their impact in various topics from physics and international relations. The results underscore the potential of information-theoretical tools to inform strategic decision-making in research funding, interdisciplinary collaborations, and journal policy planning.

In the following sections, we detail the theoretical methods, methodological implementation, and results of this study. By analyzing the geometrical properties and topology structure of knowledge networks, we contribute to advancing our understanding of how scientific knowledge evolves and how stakeholders can effectively navigate its dynamic landscape.

2. Methods used

We employ an information-theoretical metric, represented by the equation:

$$d(X,Y) = 1 - \frac{M(X,Y)}{H(X,Y)}. \tag{1}$$

This metric is derived from mutual information $M(X,Y)$ and the joint Shannon entropy $H(X,Y)$, which quantify the relationship between two stochastic variables X and Y representing the term frequencies of two distinct concepts. Given a probability $P(x_k,t) = N_c(k,t)/N(t)$ for a concept c to be cited exactly k times in a set of $N_c(t)$ documents from an $N(t)$-document corpus for the period until time t, the mutual information and the joint entropy are calculated as:

$$M(X,Y) = \sum_{k,m=0}^{\infty} P(x_k,y_m)\log_2\left(\frac{P(x_k,y_m)}{P(x_k)P(y_m)}\right), \tag{2}$$

$$H(X,Y) = \sum_{k,m=0}^{\infty} P(x_k,y_m)\log_2(P(x_k,y_m)). \tag{3}$$

Here, $P(x_k,y_k)$ and $P(x_m) = \sum_{x_k} P(x_k,y_m)$ denote the joint and marginal probability distributions, respectively.

The metric $d(X,Y)$, known as the variation of information (VI) or the shared information distance, exhibits metric properties such as the identity of indiscernibles, symmetry, and the triangle inequality[11][13][14][15]. When two concepts share an identical set of documents with the same term

frequencies, $d(X,Y) = 0$; conversely, when concepts never cooccur in any document, $d(X,Y) = 1$. In particular, including $k, m = 0$ in the definitions of $M(X,Y)$ and $H(X,Y)$ provides a common normalization constant, ensuring a consistent scale for comparing distances $d(X,Y)$ across different pairs of concepts within a document corpus of size $N(t)$ when t is fixed. However, comparing distances between concepts at different times when size of the relevant document corpus is changing necessitates an additional normalization factor:

$$\tilde{d} = \frac{d}{\log_2 N(t)}, \tag{4}$$

where $\log_2 N(t)$ is the maximal entropy that any concept may possibly reach in case of the uniform term frequency distribution across a given documents set.

The properties of d as a metric are intuitive and robust for comparison. The triangle inequality, in particular, implies that if two concepts (or clusters) are both close to a third, they cannot be too far apart from each other. This property allows us to infer potential relationships between concepts, as proximity to a common third concept suggests likely closeness. Such qualities make the VI metric a powerful tool for exploring complex relationships and predicting links within a network of concepts.

The symmetric similarity matrix, derived from pairwise distance calculations for a selected set of concepts, encapsulates information about the topological properties and complexity of the corresponding knowledge network. Our findings indicate that temporal changes in normalized or non-normalized pairwise distances, $|\Delta d/\Delta t|$, expressed in terms of matrices of a related velocity of concepts, provide additional insight into the temporal evolution of the network.

We compute the relative velocity matrix by calculating the differences between elements of the distance matrices at different time points $t_n, n \in \mathbb{Z}$. Using a constant time interval $\Delta t = t_{n+1} - t_n$, set to one year, we reduce noise and computational overhead. For simplicity, we also set $\Delta t = 1$ in the following notation. To ensure comparability across different scales of distance values, we calculate the relative speed for each element in the velocity matrix as follows:

$$\tilde{v}_{ij} = \frac{\tilde{d}_{ij}^{\,t_{n+1}} - \tilde{d}_{ij}^{\,t_n}}{\tilde{d}_{ij}^{\,t_{n+1}} + \tilde{d}_{ij}^{\,t_n}}. \tag{5}$$

As we demonstrate in the next section, the velocity matrix $\tilde{V}$ resulting from Eq. (5) contains only negative elements, indicating an increase in mutual information between concepts as new documents emerge. By computing a Minimum Spanning Tree (MST) of this velocity matrix, we can identify the

elements and patterns corresponding to the fastest and most active connections within the network of concepts at each time step. Notably, the vertex degree distribution in such spanning tree networks follows a power-law distribution, a characteristic commonly observed in social science studies and indicative of scale-free network behavior.

A potential concern with the use of the Variation of Information metric and the resulting velocity matrix is that the spanning tree method, and consequently the derived concept connectivity order, may be biased toward the most frequently occurring concepts. These commonly occurring concepts can appear as dominant elements within "hot" topics, clusters characterized by the highest negative relative velocity. This prominence may not always reflect the true structural importance of these concepts, but rather their ubiquity in the dataset. To mitigate this bias, we suggest incorporating a filtering approach based on the residual entropy framework proposed by Martini[5]. This approach allows for the identification and exclusion of concepts that contribute minimally to the reduction of uncertainty within the topic, thus enhancing the robustness of connectivity patterns and ensuring that the identified clusters reflect meaningful conceptual relationships rather than mere frequency-driven prominence.

It should be mentioned that using non-normalized distances d in Eq. (5) introduces positive matrix elements, capturing relaxation dynamics within the concept network while preserving the convergence patterns observed in the case of normalized velocity matrices. This observed relaxation suggests that, without reinforcement from new publications, certain connections may gradually weaken (become more rare) over time due to the increasing total number of documents in the corpus. By applying normalization, we remove this effect, enabling the analysis to focus solely on network compression driven by entropic forces introduced by new publications.

The effect of strengthening knowledge network interactions, driven by additional information stored in emerging publications, can be quantified through the properties of the network represented as a geometric simplicial complex. The metric $\tilde{d}$ enables us to interpret a group of $n+1$ concepts (nodes) as an n-dimensional simplex with a well-defined volume[16]:

$$\tilde{V}_n^2 = \frac{(-1)^{n+1}}{2^n (n!)^2} \det \begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & 0 & \tilde{d}_{12}^2 & \dots & \tilde{d}_{1(n+1)}^2 \\ 1 & \tilde{d}_{12}^2 & 0 & \dots & \tilde{d}_{2(n+1)}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \tilde{d}_{1(n+1)}^2 & \tilde{d}_{2(n+1)}^2 & \dots & 0 \end{pmatrix}. \tag{6}$$

We can quantify the high-dimensional (i.e., high-order) impact of different sources (aggregated or individual) on the structure of the knowledge network as a change in the volume of a respective simplex.

To evaluate the impact of a particular knowledge broker-related subset of documents on the topology of the knowledge network formed from a group of n concepts, we calculate the normalized relative difference of the topic volume (see Appendix B):

$$\frac{\Delta\tilde{V}_n}{N_r} = \frac{\tilde{V}_n^f - \tilde{V}_n^a}{N_r \tilde{V}_n^a}, \tag{7}$$

where $\tilde{V}_n^f > \tilde{V}_n^a$, $\tilde{V}_n^a$ is the topic volume obtained from all documents, and $\tilde{V}_n^f$ is the volume derived from a filtered subset of documents that excludes publications associated with specific knowledge brokers identified through document metadata. The resulting difference is normalized by the number of relevant documents N_r associated with a particular broker, providing a measure of its impact per document.

A document is associated with a broker if it contains at least two concepts from the topic considered. When $\Delta\tilde{V}_n/N_r$ approaches zero, the agent's impact is negligible, indicating that the volumes of filtered and unfiltered topics are nearly identical. In contrast, a higher value of $\Delta\tilde{V}_n/N_r$ reflects a stronger influence of the broker, as the filtered volume $\tilde{V}_n^f$ is significantly larger than the unfiltered volume $\tilde{V}_n^a$. This increase suggests that the structure of the knowledge network undergoes noticeable compression due to the presence of publications associated with the broker. Drawing an analogy from physics, each knowledge broker (such as an author, journal, publisher, or institution) acts as an *agent* that performs conceptual "work" by increasing the density and connectivity of the knowledge network.

In the following section, we first analyze the dynamics of the topic network by calculating various parameters of normalized and non-normalized velocity matrices and examining the evolution of network volume as a function of emerging documents. Second, we assess the structure and quantify the impact of knowledge brokers, such as journals and countries, as "thermodynamic" agents using the normalized relative volume measure from Eq. (7). Examples from topics in physics and international relations illustrate these analyses.

2.1. Dataset description

For our analysis of physics-related topics, we used 421,524 research papers from the high-energy physics and astronomy sections of the ArXiv preprint server (http://arxiv.org) covering the period 1990–2018. For international relations topics, we analyzed 10,370 documents from the JSTOR platform (http://constellate.org) published between 2010 and 2024, focusing on issues related to international security. For the ArXiv dataset, the concept frequency data was obtained from the ScienceWise platform (now ProphyScience, https://www.prophy.science/)[17]. All JSTOR documents were preprocessed using an Extract, Load, and Transform (ELT) pipeline in Google BigQuery to load keyword metadata from document sections and match the cleaned collection of keywords with n-gram frequencies (uni-, bi-, and trigrams) originally available in the JSTOR database.

We acknowledge that the datasets used in this study focus exclusively on English-language documents, which may result in an incomplete representation of the full spectrum of research output in the fields of physics and international relations. This language bias could potentially exclude valuable contributions published in other languages, thus limiting the generalizability of our findings across global research communities.

To identify various knowledge brokers, the metadata for publications in JSTOR and ArXiv were enriched, where possible, with information from the Web of Science (WOS) database by matching digital object identifiers (DOI) and publication titles if the DOI is not available. Due to differences in document coverage and metadata between JSTOR, ArXiv, and WOS, accurate matches were achieved for only a fraction of the available documents. For ArXiv publications, 54.2% of records were successfully matched, while for JSTOR, the match rate reached 97.51%. This enrichment allowed the identification of brokers, such as journals, publishers, and countries, facilitating a comprehensive assessment of their impact on the structure and dynamics of knowledge networks.

3. Results

Developing a carefully constructed scientific ontology as a comprehensive collection of scientific concepts is crucial for keyword-frequency text analysis[18][19]. This discussion will not explore the methods and challenges of creating such a dictionary; for a detailed review on this topic, we recommend consulting[20]. However, once established, this scientific dictionary enables the construction and analysis of a complete distance matrix encompassing all known scientific concepts.

However, recurrently recalculating a large-distance matrix to extract network dynamics across the entire scientific domain proves computationally inefficient. In addition, scientific ontologies continually evolve with the introduction of new concepts that define novel models, methods, or experiments. Consequently, regular updates to existing distance matrices become essential.

One possible approach to reduce computation time is to select a subset of concepts closely related to a specific seed concept, which we will refer to as a 'topic.' This approach significantly reduces computational scope, since the size of the relevant distance matrix is determined by the degree centrality of the seed concept in the entire network, typically much smaller than the total number of nodes (concepts). In practice, the collection of concepts related to a topic can be obtained using the following filtering algorithm.

Algorithm 1 Pseudocode for topic selection algorithm

Input: Seed concept, term frequency threshold $k > 1$, document set D, frequency threshold $n\%$, cut-off distance b
Output: Relevant set of concepts within the topic
Step 1: Define Document Set
for each document d_i in the corpus **do**
  if Seed concept appears in d_i with term frequency $\text{tf} = k > 1$ **then**
    Add d_i to document set D
  end if
end for
Step 2: Select Relevant Concepts
for each distinct concept c_j in document set D **do**
  if Concept c_j appears in at least $n\%$ of documents in D **then**
    Add c_j to concept set C
  end if
end for
Step 3: Filter by Distance
for each concept c_j in concept set C **do**
  if Distance $d(\text{Seed concept}, c_j) < b$ **then**
    Retain c_j in the final set of relevant concepts
  end if
end for
Return Set of relevant concepts within the topic

The typical number of relevant concepts obtained from the proposed algorithm when $n = 5\%$ and $b = 0.98$ for the ArXiv and JSTOR collections is on the order of hundreds. By changing these parameters, we can increase or decrease the number of concepts in a topic relevant to the chosen seed concept. For example, following this algorithm when the parameters are set to mentioned values, we extract 268 related concepts for the "Reheating" seed concept associated with cosmic inflation at 2018.[1] Calculating a single velocity matrix under these parameters using a `c2-standard-16` machine type in Google Cloud Compute Engine takes less than an hour. The calculations were performed with Wolfram Mathematica 13.2 on Linux, while GCP Cloud MySQL 5.7 was used to host the database.

While the analysis of a larger collection of concepts provides a more detailed and nuanced evaluation of the topic under investigation, the calculation of larger distance matrices becomes computationally prohibitive, as the associated complexity grows exponentially with the number of matrix elements. This challenge is further compounded by the structure of the framework itself, as the expression used to measure the impact of knowledge brokers on the knowledge network structure depends on the number of nodes n in the network, scaling as a^{n+1}, where the constant a is typically close to one (see Appendix B for detailed derivations). This constant a is related to the ratio of the logarithms of the number of documents that characterize the data set and the contributions of the knowledge broker. As the number of concepts n increases, the impact calculation becomes biased toward brokers represented by larger volumes of documents, effectively narrowing the range of identifiable brokers to those with substantial representation in the dataset.

Moreover, increasing the size of the concept set may inadvertently obscure the influence of smaller brokers, as their contributions become statistically insignificant relative to the overall dataset. This raises the risk of overlooking potentially important, albeit less prolific, agents that may play a crucial role in niche subfields or emerging research areas. To mitigate these limitations, a balanced approach is needed: one that carefully selects the most relevant concepts while ensuring that the computational complexity remains tractable. Potential strategies include partitioning the knowledge network into smaller, thematically coherent sub-networks or employing dimensionality reduction techniques to preserve the core structure of the network without overwhelming computational resources. These approaches could help maintain sensitivity to both prominent and less-represented knowledge brokers, ensuring a more comprehensive analysis of their contributions to the evolution of the knowledge network.

In Figure 1, we present the semantic landscape of the related topic by illustrating a 2D Delaunay mesh of a corresponding graph[21] at different moments of its evolution.

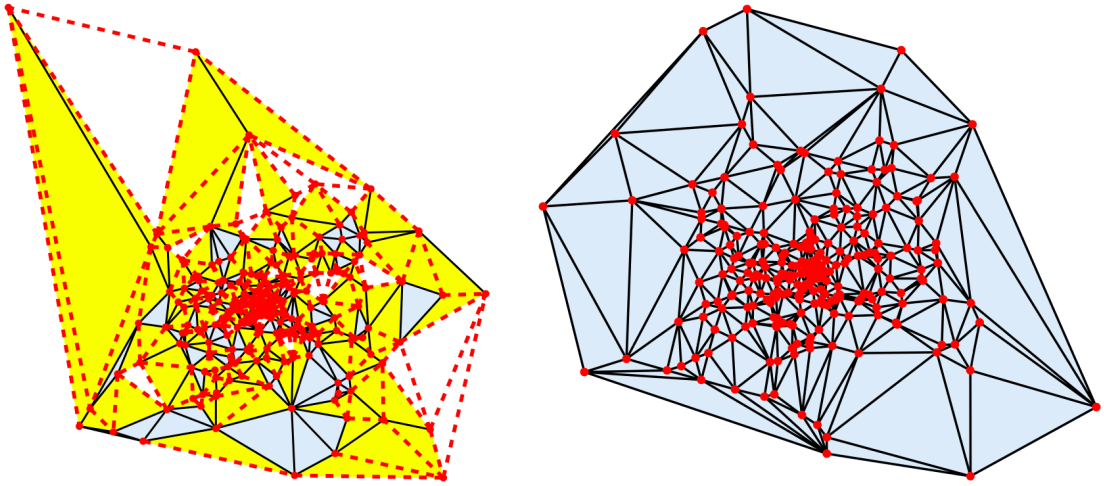

Figure 1. Two-dimensional Delaunay mesh of the normalized distance matrix of a "Reheating" topic, calculated for 2002 (left) and 2018 (right). Blue triangles represent areas where all three vertices (concepts) are connected by edges with distances below the specified normalized cut-off distance $\tilde{a} = 0.939042$ (solid black lines). Yellow triangles indicate areas where at least one distance exceeds the cut-off (dashed red lines). White triangles correspond to areas where all edge distances are above the cut-off.

Figure 1 illustrates a typical process of topic evolution during ongoing knowledge production, where initially present knowledge gaps are progressively filled with information from new publications. It is important to note that knowledge production and the increasing density of a topic do not cease once all concepts become strongly connected (i.e., when distances between them fall below a given threshold). Instead, the network continues to evolve as new concepts are continually attracted to the core of the topic, further increasing its density and expanding its conceptual boundaries.

This dynamic nature of knowledge networks underscores the continuous growth and enrichment of scientific discourse. The proposed algorithm for selecting relevant concepts captures this progression by identifying different—and increasingly larger—sets of relevant concepts when applied to database snapshots from different time periods. As the time frame expands and new data are incorporated, the conceptual landscape becomes both denser and more extensive, reflecting the cumulative nature of knowledge production and its impact on the structure of the topic.

In this study, we discover patterns of topic evolution that highlight how concepts become integrated into the topic's core structure and how their relationships evolve over time. These patterns may pave

the way for formulating more robust methods of link prediction, where the strength and formation of new conceptual connections can be anticipated based on historical trends. Understanding these principles can enable the development of predictive models capable of forecasting future link weights and emerging connections. Such methods could be instrumental for strategic research planning, helping stakeholders identify critical conceptual gaps or future collaboration opportunities that may shape the next phase of scientific discovery.

Analyzing the network dynamics associated with the non-normalized velocity matrix for the "Reheating" topic, we observe a typical pattern of vertex order dynamics that remains consistent across multiple topics. Figure 2 presents the vertex order distributions for the positive and negative parts of the velocity matrix, calculated using Eq. (5) for non-normalized distances *d*.

In this context, the order of the vertex reflects the *centrality rate* of a concept, indicating how its mutual information with other concepts in the network changes, whether it is increasing (convergence) or decreasing (relaxation). As illustrated in Figure 2(left), converging subnets tend to have lower vertex orders and, consequently, lower connectivity compared to their relaxation counterparts. This observation underscores the localized nature of topic updates from the literature, where concept relationships are strengthened or weakened based on the arrival of new information. In particular, each concept on the topic simultaneously belongs to both the converging and relaxation subnets, with its centrality rate adjusting in response to signals from new publications.

Since Algorithm 1 fixes the number of concepts within the studied topic, the number of possible links between concepts remains constant. This constraint results in an almost symmetric distribution of vertex orders in the converging and divergent components of the velocity matrix at any given time.

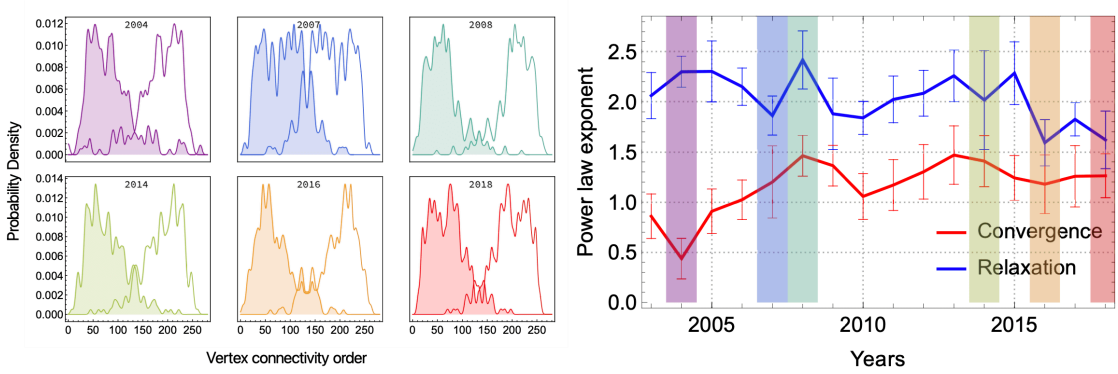

Figure 2. Kernel-smoothed vertex order probability distribution for the positive and negative components of the velocity matrix associated with the evolution of the "Reheating" topic (left) consists of 268 concepts. Evolution of the power-law exponent for vertex order distributions in the spanning tree network of the convergent (negative) and divergent (positive) components of the velocity matrix (right). The same colors in both plots indicate the same years.

In this context, analyzing the spanning trees of the converging and diverging subnets provides valuable information on the dynamics of mutual information exchange between concepts over a given period. The central nodes in these trees correspond to concepts that experience the most significant gains or losses in mutual information. Their neighboring nodes reveal trends within the topic, identifying concepts that either attract the most attention or, conversely, become less frequently referenced during this period. Our analysis shows that the distribution of vertex connectivity orders in both subnet trees follows a power law pattern $k^{-\alpha}$ (see Appendix A for details), with exponents α differing between the subnetworks and evolving over time.

A decrease in this exponent indicates that concepts with higher connectivity orders become more likely, and the corresponding network gains more hubs (nodes with many connections). In contrast, a higher value of α implies a steeper distribution, where highly connected nodes are rare and most nodes have low connectivity. It is important to note that changes in the power-law exponent correspond to different processes in convergent and divergent networks. In the case of a convergent network, a decrease in α corresponds to a growth in topic popularity, where publications increasingly refer to a few core concepts. Conversely, a decrease in α in a divergent network suggests that concepts within the topic are becoming more dispersed. This may indicate that the topic is becoming "colder"

or more fragmented, as peripheral or unrelated references contribute to a reduction in the topic's overall cohesion.

The dynamics of the power-law exponent, as shown in Figure 2 (right), serve as an informative parameter for describing the properties of the convergent and divergent subnets. As observed, the power-law exponent for the divergent subnetwork of the velocity matrix is typically higher than for its convergent counterpart. This indicates that the divergent network is more dispersed with fewer hubs. In contrast, the convergent subnetwork, with a smaller α, demonstrates that discussions in the literature that affect knowledge dynamics are centered around a few core concepts with close connections to other concepts. The "hotter" the discussion, the higher the connectivity within these hubs of concepts.

It is important to note that the composition of both networks, in terms of the set of concepts, remains unchanged, as the set of concepts used to construct both networks stays constant. What changes is the connectivity between the concepts. A set of concepts that initially had high connectivity in the convergent network may, at the next point in time, be replaced by a completely different set of highly connected concepts. This is especially true for the convergent subnet, as the number of interacting concepts that is related to the vertex connectivity order, as shown in the Figure 2 (left), is smaller than the total number of concepts associated with the topic. The dynamics of the exponent in the power-law distribution, particularly for the convergent network, thus determine the intensity of connections between various concepts within the topic and reflect the changes in the average connectivity over time.

The velocity matrix calculated using normalized VI distances contains only negative elements at any time, indicating the continuous convergence of distances in terms of normalized metrics within the corresponding knowledge network as the number of emerging documents increases. Each node in the network associated with the normalized velocity matrix has a connectivity order equal to the total number of nodes in the topic, resulting in a complete graph for this network. However, the distribution of values for the relative velocities within the normalized matrix is not uniform. As shown in Figure 3 (left), the velocity distribution in the early stages of topic development is broader compared to later stages. Over time, the relative velocities between concepts decrease and eventually become nearly the same across the topic.

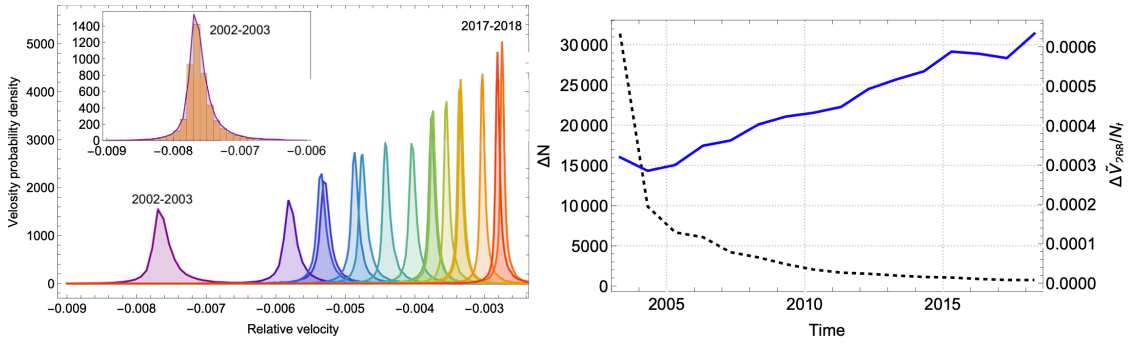

Figure 3. The dynamics of the non-zero relative velocity distribution in the 'Reheating' knowledge network topic (left), calculated with a one-year time step. The dynamics in the topic's relative volume per single publication calculated over time with the year time step (black dashed line) from new ΔN documents published every year (blue line) in ArXiv on this topic.

Observed in Figure 3, the behavior reflects the natural evolution of a topic as new publications gradually reduce the variability of conceptual relationships, stabilizing the structure of the knowledge network. In the early stages, when the topic is less developed (i.e., consists of fewer concepts), introducing new concepts and connections between them leads to significant adjustments in the network geometry, resulting in higher velocities and a broader relative velocity distribution. As the topic matures, the most relevant connections are established, and the network exhibits slower and more uniform changes that continuously compress the size of the topic. This trend highlights the self-organizing nature of knowledge networks, where initial dynamism gives way to a stable structure as the topic approaches a state of "equilibrium."

Defining the topic volume as the volume of an n-dimensional simplex allows for a comprehensive assessment of the impact of selected publications on the geometry of the studied topic, considering all possible connections between concepts simultaneously. From Figure 3 (right), we observe that the relative normalized change in the topic volume, calculated using Eq.(7), decreases over time as new documents emerge. The rate of change for this parameter, like the relative velocities of the concepts, decreases with time, even as the number of relevant documents N_r increases each year.

The observed behavior of the topic volume as a function of the fraction of participating documents allows us to develop a method to evaluate the impact of publications that share certain metadata, such as first authors, author affiliations, or institutional associations. The impact value can be derived from the analysis of a generalized "velocity" matrix-like approach, where we trace not temporal changes

but the effect that a specific group of documents has on the structure of the studied topic. Document metadata enables broad classifications, and our approach quantifies the impact of a generalized class of documents associated with an abstract knowledge broker or "agent" that shape the topic. This agent, as we mentioned above, can represent an author, journal, publisher, grant agency, institution, country, geographic region, or something else based on available information. Additionally, our method allows us to take a more abstract approach in identifying the influence of one topic on another, provided that we can identify relevant groups of documents in the data set.

In Figures 4 and 5, we present the relative volume changes of the topics "Reheating" and "International Security" attributed to documents associated with different countries and publishing companies (journals for physics topics). To calculate the impact at the country level, we classified the ArXiv documents by the affiliation of the corresponding author recorded in the Web of Science database, and computed the relative volume using Eq. (7). However, the obtained document-country associations do not account for coauthor affiliations, which means that our results, given as an example, are limited to the primary affiliations of the corresponding authors. This limitation highlights an avenue for future refinement of the method by incorporating a broader range of metadata to capture a more comprehensive picture of the influence of documents.

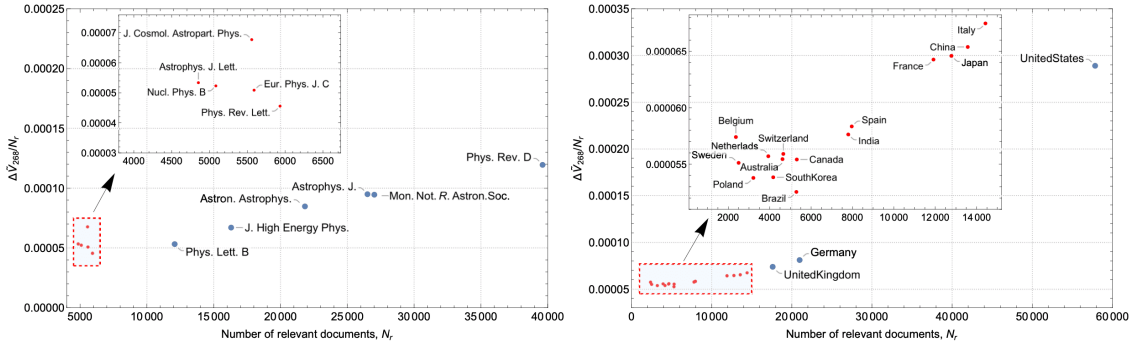

Figure 4. The impact on the 'Reheating' topic knowledge network with $n = 268$ concepts from journals (left) and countries (right), measured as the normalized relative change in the topic's simplex volume based on the number of relevant documents N_r published in ArXiv for the period 1990–2018.

We do not face the mentioned ambiguity in the case of publisher- or journal-related collections, as these associations are more straightforward and unambiguous. Each document is uniquely attributed to a specific journal or publisher, allowing for a precise analysis of their influence on the topic's

doi.org/10.32388/O5NMBG.2

structure. This direct correspondence ensures that the calculated normalized relative volume changes genuinely reflect the impact of these publishing entities without the need to account for overlapping contributions or multiple affiliations.

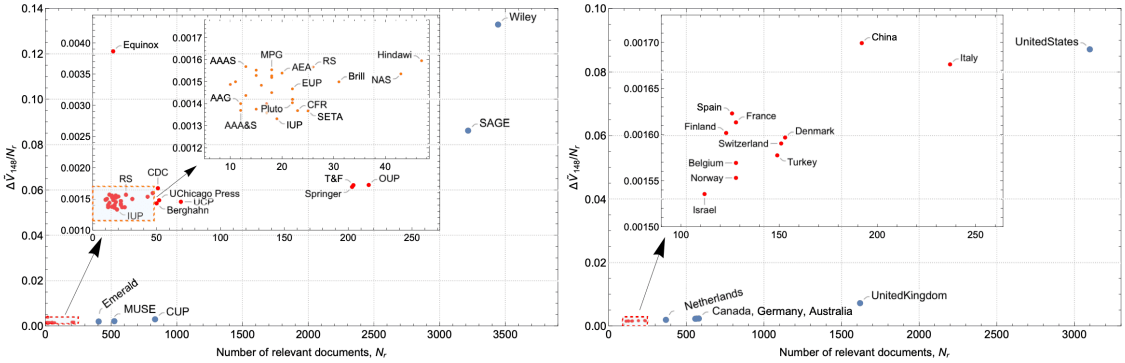

Figure 5. The value of the impact on the "International security" topic knowledge network with $n = 148$ concepts from documents related to countries (right) and publishers (left) measured as relative topic volume change per relevant document published in the period 2010-2024.

Analysis of the structure of publisher and journal impact can provide valuable insights into how the dissemination of knowledge through specific channels shapes the evolution of scientific topics. For example, it can reveal whether certain publishers or journals specialize in fostering particular research areas or play a significant role in diversifying related concepts within a topic. These findings improve our understanding of the academic ecosystem and help identify key contributors to the growth and development of specific domains.

The impact on the geometry of a topic made by an external agent can be analyzed using a similar approach to that employed to quantify dynamic changes in the corresponding network topology. By calculating the difference between the distance matrices - one derived from all relevant documents and another from a subset that excludes publications associated with a specific agent - we obtain a generalized velocity matrix. This matrix allows us to extract both the overall structure of the impact and its most significant contributions.

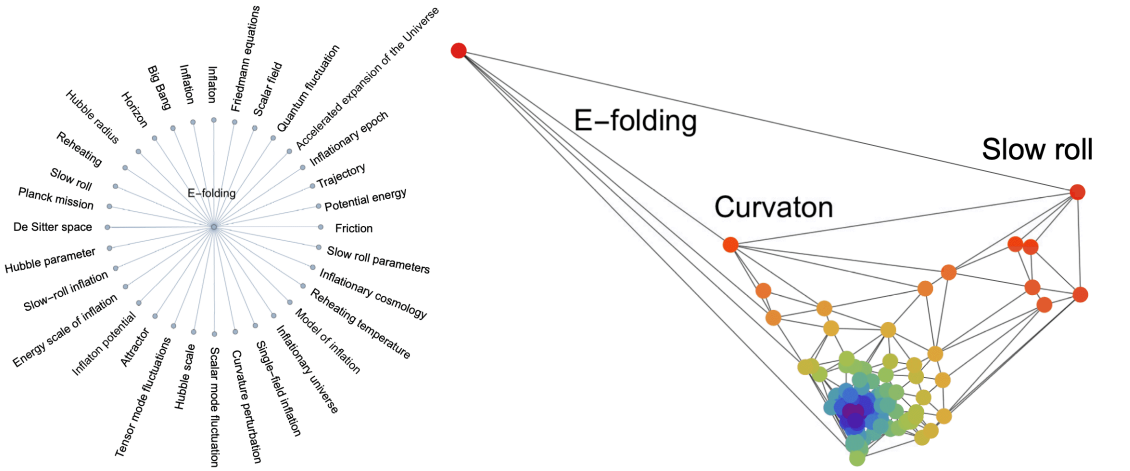

Figure 6. The structure of the largest impact made on "Reheating" topic network structure from documents whose corresponding author has affiliation in the institutions from United States of America. The node of the highest connectivity and its related vertices, extracted from the graph difference spanning tree network (left) and Delaunay complex for this network (right).

As an example, Figure 6 illustrates the largest contribution to the topic "Reheating" from publications categorized under "United States" (i.e., where the corresponding author is affiliated with an institution in the United States of America). The subgraph for the corresponding topic highlights this contribution. Additionally, the Delaunay complex computed over the spanning tree of the generalized velocity matrix provides a comprehensive view of changes in the information topology of the studied topic. In Figure 6 (right), the multidimensional structure of the contribution and the relative magnitude of the changes are depicted. In particular, the triangle formed by the concepts "E-folding"[2], "Slow roll"[3], and "Curvaton"[4] represents the most impactful physics concepts within the "United States" document category during the entire period from 1990 to 2018.

Analyzing the structure of the impact on a country level in case of an international security topic reveals the most important concepts that shape the topic's knowledge geometry structure. In Figure 7 we present cloud maps of concepts that reflect the structure of the impact from two the most impactful countries, the United States and the United Kingdom. The concepts with the highest module of a generalized velocity, i.e. those having the largest order in the corresponding spanning-tree network, are the most influential in determining the dynamics of the topic's evolution. These concepts act as "hubs" within the knowledge network, serving as critical points for knowledge aggregation or

dispersion. The higher the generalized velocity of a concept, the more actively it participates in reshaping the network's structure, either by reinforcing existing connections or by introducing new conceptual pathways.

Figure 7. The structure of the international security topic dynamics from publications in the JSTOR database associated with the United States (left) and the United Kingdom (right).

In the case of the United States, the predominant influence stems from concepts such as "Activists," "Harvard University," "Columbia," "Princeton University", "President", "Stanford University" and "Immigrants." These concepts highlight the country's significant contribution to the academic and institutional discourse surrounding international security. The presence of prominent universities and terms related to social movements and migration indicates a focus on thought leadership, policy influence, and the social dimensions of security. This suggests that the impact of the United States is characterized by both institutional authority and grassroots activism, reflecting a multidimensional approach to shaping the knowledge landscape of international security.

In contrast, the contributions of the United Kingdom are defined by concepts such as "Global Governance", "International Affairs", "International Politics," "Terrorism," "International Security," "Mapping" and "Insecurity". These terms underscore the UK's focus on geopolitical strategy, regulatory frameworks, and global cooperation. The emphasis on "Global Governance" and "International Security" illustrates the UK's role in fostering international collaborations and addressing global security challenges. Meanwhile, the inclusion of "Mapping" and "Insecurity" points

to analytical approaches and concerns over instability, suggesting a more methodical and systemic contribution to the discourse on international security.

This comparison highlights how country-level contributions not only introduce distinct thematic priorities but also shape the overall conceptual structure of the topic. The United States' influence appears to center on academic and activist-driven discussions, whereas the United Kingdom's contributions emphasize governance, policy-making, and the analytical mapping of global security dynamics.

The cloud maps illustrate not only the prominence of these concepts, but also the breadth of their influence within the topic. Concepts with higher generalized velocities often form tightly interconnected clusters, indicating areas of active discourse where new knowledge is rapidly integrated. In contrast, peripheral concepts with lower velocities represent niche discussions or emerging areas that may later become more central as the topic evolves. This analysis underscores how country-level contributions can shape the overall "knowledge topology" of international security by reinforcing dominant narratives or introducing alternative perspectives.

4. Conclusions and outlook

This study presents a novel framework for analyzing the dynamics of scientific topics and their evolution within knowledge networks using information-theoretical metrics. Using the variation of information (VI) metric and normalized velocity matrices, the methodology captures temporal and structural changes in concept networks, providing insight into the mechanisms driving knowledge production and dissemination. The approach provides a robust tool for quantifying the influence of various knowledge brokers – such as countries, journals, institutions, authors, grant organizations, etc. – on the geometry of these networks. In this research, we show examples of measuring contributions from the documents selected by the first author affiliations country and article publishing organization in the case of topics from physics and international relations.

The key findings of this research include:

- Dynamic Knowledge Compression: The analysis reveals that as topics evolve, knowledge networks exhibit self-organizing behavior, transitioning from high variability in conceptual relationships to a stable structure characterized by reduced topic volume. This pattern reflects the natural convergence of scientific concepts as new publications emerge.

- Agent-Specific Impacts: By associating changes in topic volume with specific knowledge brokers or agents, the methodology quantifies the influence of these entities on shaping topic dynamics. This capability highlights the role and shows the structure of prominent contributors and dissemination channels in the advancement of specific research areas.
- Scalability and Adaptability: The approach demonstrates its effectiveness across diverse datasets, from physics-related topics based on ArXiv publications to international relations topics extracted from JSTOR-indexed publications, showcasing its adaptability to different disciplines and research landscapes.

The practical implications of these findings extend to academic journals, funding agencies, and research institutions, enabling data-driven decision-making in resource allocation, trend prediction, and fostering interdisciplinary collaborations. The use of normalized geometric metrics also allows for cross-comparison of topic dynamics across databases, providing a standardized measure of agent impact across different datasets.

Looking ahead, future research should address the universality of these findings in diverse datasets and disciplines. Incorporating additional methods for the extraction of new concepts and expanding metadata analysis could further refine the accuracy and applicability of the proposed framework. Furthermore, the integration of dynamic clustering algorithms and advanced visualization techniques could enhance the interpretability of knowledge network structures and their evolution. Furthermore, analysis of the dynamics of the topic distance matrix suggests that various link prediction methods could be applied in future research to gain valuable insights into how knowledge gaps are bridged between different disciplines. These methods have the potential to reveal the role of knowledge brokers in facilitating these connections, providing a deeper understanding of the processes that drive interdisciplinary integration and knowledge dissemination.

Available information on topic distance matrix dynamics allows for the applicability of the various link prediction methods that in this case will give a valuable incite about knowledge gaps bridging patterns in various disciplines and role of different knowledge brokers in this process.

By establishing a quantitative framework for understanding and navigating the complex landscape of scientific discourse, this study contributes to the advancement of the field of knowledge network analysis and strategic planning in science policy and research management.

Open science practices

Current research utilizes JSTOR and ArXiv (http://arxiv.org), both publicly accessible sources of scientific metadata and document texts. In JSTOR (https://constellate.org/), the text of the documents is parsed and uni-, bi-, and trigrams are already extracted, allowing subsequent matching of terms with existing dictionaries to compile the frequencies of ontology terms. The dataset and the code for analyzing the JSTOR corpus are available on Zenodo (DOI: 10.5281/zenodo.14638434). For scientific dictionary-based analysis, the study uses the dictionary of concepts in high-energy physics and astronomy provided by the ScienceWise platform. However, access to the ScienceWise dictionary is restricted due to confidentiality agreements established with the platform.

Appendix A: Vertex connectivity order distributions in the velocity matrix

This study introduces the velocity matrix, a tool for quantifying temporal changes in the topic distance matrix, offering insights into the dynamic evolution of conceptual relationships within a topic network. Calculated from non-normalized distances, this matrix includes positive and negative elements that reflect changes in interconcept distances over time intervals. Positive values indicate increasing distances, suggesting a weakening of mutual relationships or divergence in the topic's structure, while negative values signify decreasing distances, implying strengthening relationships or convergence.

By separately analyzing the positive and negative components of the velocity matrix, concepts with the highest relative velocity in divergent and convergent subnets can be identified. Applying a minimum-spanning tree algorithm to the weighted adjacency matrix of the negative velocity matrix highlights the most interconnected and rapidly converging concepts. These concepts represent key sources of information context or the main spreaders within the topic network. Similarly, analysis of the weighted adjacency matrix of the positive velocity matrix using the maximum spanning tree algorithm reveals the concepts that are most divergent, representing areas where conceptual relationships are weakening or fragmenting. These divergent concepts may correspond to emerging subtopics, shifting research priorities, or outliers that challenge established paradigms within the topic network.

Vertex orders in converging and divergent spanning tree networks are distributed according to a power law $k^{-\alpha}$, indicating the presence of scale-free network structures. The value of the α exponent can be interpreted as the inverse "temperature" of the corresponding network. A lower value of α corresponds to a heavier tail in the distribution, indicating that certain concepts exhibit high connectivity. This high connectivity may arise from their intensive usage or, conversely, from the less frequent usage of these concepts in the literature.

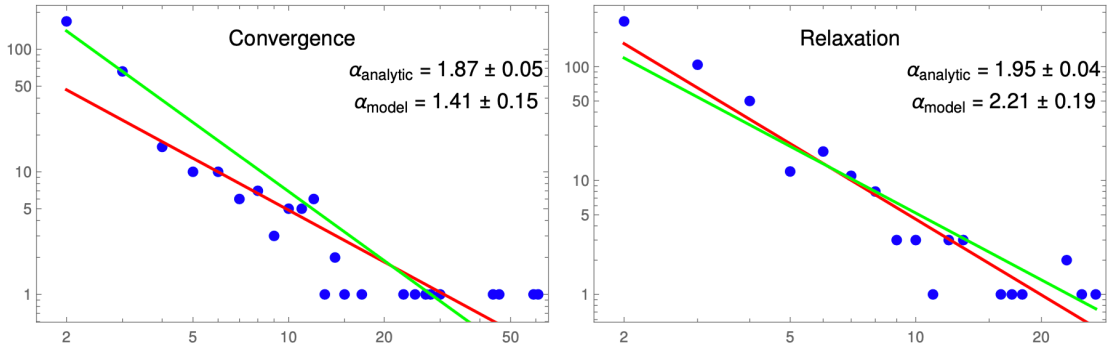

Figure A1. Log-log plot for vertex orders within the "Reheating" related spanning tree network of the positive and negative parts of the velocity matrix for 2017-2018. The x-axis represents the vertex order (degree), and the y-axis shows the number of nodes. The vertex order distribution follows a power-law $k^{-\alpha}$. The exponents were obtained from linear regression (α_{analytic} - red line) and model-derived[22] (α_{model} - green line).

The presence of positive elements in the velocity matrix is primarily an artifact of the growth of the data set, as the increasing number of documents inflates the entropy and distances between concepts. To mitigate this effect, normalization of the topic distance matrix with respect to the dataset's maximum entropy, $\log_2 N$ (where N is the number of documents), ensures size invariance. This normalization removes positive elements, isolating genuine convergence patterns and revealing conceptual compression driven by new publications and strengthened relationships.

Appendix B: The Relative Simplex Volume

To quantify the impact of specific collections of documents associated with certain agents on the geometry of a knowledge network, we calculate the normalized relative change of a topic volume $\Delta\tilde{V}_n/N_r$. This measure provides a robust comparison of topic volumes derived from full and filtered

document corpora, accounting for differences in dataset sizes and preserving the network's structural properties. The relative change of a topic volume of a dimension n is defined as a relative difference:

$$\Delta\tilde{V}_n = \frac{\tilde{V}_n^f - \tilde{V}_n^a}{\tilde{V}_n^a}, \tag{B1}$$

where $\tilde{V}_a$ and $\tilde{V}_f$ are the simplex volumes of the network of n concepts calculated from the full and filtered document sets, respectively, containing N_a and N_f total documents. In a filtered data set, we leave only those documents that are not related to the specific agent (e.g., country, journal, author, etc.). Volumes $\tilde{V}_a$ and $\tilde{V}_f$ are calculated based on normalized distances $\tilde{d} = d/\log_2 N$, where N represents the size of the actual set of documents[11]. This normalization accounts for the size of the data set and ensures consistency in comparing network volumes across data sets.

The normalized relative volume change is calculated using Eqs. (6) and (B1) as:

$$\frac{\Delta\tilde{V}_n}{N_r} = \frac{1}{N_r}\frac{\sqrt{\frac{\det B_f}{(\log_2 N_f)^{2(n+1)}}} - \sqrt{\frac{\det B_a}{(\log_2 N_a)^{2(n+1)}}}}{\sqrt{\frac{\det B_a}{(\log_2 N_a)^{2(n+1)}}}} = \frac{1}{N_r}\left(\left(\frac{\log_2 N_a}{\log_2 N_f}\right)^{n+1}\frac{\sqrt{\det B_f}}{\sqrt{\det B_a}} - 1\right), \tag{B2}$$

where $N_r = N_a - N_f$ is a number of relevant documents related to the considered agent and topic, and B_f and B_a are $(n+1)\times(n+1)$ Cayley-Menger matrices. The large sizes of the data sets and, consequently, the values of $\log_2 N_a$ or $\log_2 N_f$ of the distance normalization factors create challenges for the numerical calculations of the Cayley-Menger determinants due to the very small values of the matrix elements. To address this, we used the property $\det(cB) = c^{(n+1)}\det(B)$ to extract the normalization factors $c = 1/\log_2^2 N$ from the determinant in Eq. (6).

Expression in Eq. (B2) provides a finite quantitative measure of the geometric compression of a topic due to the influence of a selected subset of publications. Normalization by the number N_r of documents in such a subset allows us to compare the impact of agents of different sizes and measure the average creativity of the production of associated knowledge. By capturing these effects, it offers a generalized perspective on estimating how scientific publications shape the structure of the knowledge network.

One of the primary limitations of the proposed approach lies in its computational complexity. As the dimensionality n of the topic increases, the size of the distance matrix increases quadratically, leading to an exponential increase in the complexity of calculating the Cayley-Menger determinants. Furthermore, the normalized relative volume change depends on the number of concepts n such as

a^{n+1}, where a is a factor related to the ratio of logarithms in Eq. B2. This exponential dependence can significantly reduce the range of detectable agents, particularly for large topics, as only agents with a substantial number of relevant documents can meaningfully influence the network structure.

For small datasets n, the metric is less sensitive to differences in dataset size, making it more suitable for analyzing compact and well-defined topics. However, for large datasets n, the metric becomes increasingly sensitive to small differences between the full and filtered datasets, amplifying the impact of dataset size discrepancies. This sensitivity can provide valuable insights into the dynamics of highly interconnected topics but may also introduce challenges in interpreting the results for broad or diffuse topics. To address these issues, future implementations could incorporate concept filtering method based on their novelty in a considered dataset such as that introduced by[5] and extended in the author's previous publications[6].

The sensitivity of Eq.(B2) to topic size can be evaluated through numerical dimensional analysis, as illustrated in Figure B1. This analysis uses country-level data for the international security topic to explore how the normalized relative volume change depends on the number of concepts n within a topic. To investigate this relationship, we randomly select subsets of concepts from the original topic to systematically reduce its dimension, focusing on the five countries of greatest impact: the United States, the United Kingdom, Australia, Germany, and Canada.

Our findings reveal that for agents associated with a small number of relevant documents, the relationship between topic dimension and normalized agent efficiency is approximately linear. In contrast, for agents with a large number of relevant documents comparable to the size of the entire dataset, this relationship becomes nonlinear, as illustrated in Figure B1.

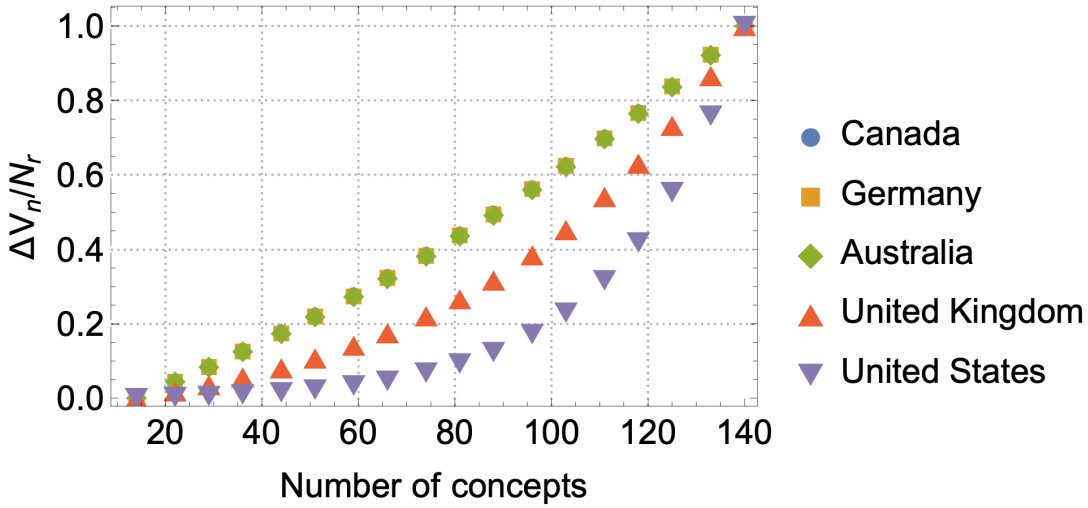

Figure B1. The rescaled dependence of country-level agent efficiency on the topic dimension n within the international security topic.

By carefully selecting the dimensionality n and applying the method to appropriately scoped topics, researchers can maximize the utility of the proposed approach while mitigating its computational limitations. This balance ensures that the method remains both practical and insightful, offering a versatile framework for exploring the impact of agents on the geometry of knowledge networks.

Appendix C: Table of Publisher Contributions

In this section, we present a comprehensive table that showcases the contributions of various publishers to the debate on international security captured in the JSTOR database for the period 2010–2024. The table includes the abbreviation and full name of each publisher, along with the number of relevant documents (N_r) associated with each publisher and their corresponding impact factor. The impact factor is calculated as the normalized relative topic volume change, reflecting the influence of each publisher's publications on the structure of the knowledge network.

This table provides insights into the extent to which different publishers contribute to shaping the conceptual landscape of international security by quantifying their relative impact on the knowledge geometry of the topic. Publishers with a higher impact factor demonstrate a stronger influence in shaping topic evolution, highlighting their pivotal role in the dissemination of critical research.

The following table summarizes these details:

doi.org/10.32388/O5NMBG.2

Journal Rank	Short Name	Full Name	N_r	Impact Value
1	Wiley	John Wiley & Sons, Inc.	3445	0.132908
2	SAGE	SAGE Publications	3217	0.086142
3	CUP	Cambridge University Press	834	0.00294469
4	MUSE	Project MUSE	524	0.00211224
5	Emerald	Emerald Group Publishing	405	0.00197213
6	OUP	Oxford University Press	216	0.00172304
7	T&F	Taylor & Francis, Ltd.	204	0.00171996
8	Springer	Springer	203	0.00169188
9	UCP	University of California Press	69	0.00145353
10	UChicago Press	The University of Chicago Press	52	0.0014776
11	CDC	Centers for Disease Control & Prevention (CDC)	51	0.00167111
12	Berghahn	Berghahn	50	0.00142858
13	Hindawi	Hindawi	47	0.00159471
14	NAS	National Academy of Sciences	43	0.00153501
15	Brill	Brill	31	0.00149908
16	RS	The Royal Society	26	0.00156647
17	SETA	SET VAKFI İktisadi İşletmesi, SETA VAKFI	25	0.00136776
18	CFR	Council on Foreign Relations	23	0.0013683
19	EUP	Edinburgh University Press	22	0.00146695
20	UIP	University of Illinois Press	22	0.00141933
21	Pluto	Pluto Journals	22	0.00140414
22	AEA	American Economic Association	20	0.0015386
23	IUP	Indiana University Press	19	0.00133152

Journal Rank	Short Name	Full Name	N_r	Impact Value
24	MPG	Max-Planck-Gesellschaft zur Foerderung der Wissenschaften	18	0.00155373
25	BioOne	BioOne	18	0.00152604
26	BMJ	BMJ	18	0.00151885
27	AR	Annual Reviews	18	0.00145007
28	CRS-York	Centre for Refugee Studies, York University	17	0.00139629
29	ASSAf	Academy of Science of South Africa	17	0.00139993
30	JBPC	John Benjamins Publishing Company	17	0.0013587
31	GESIS	GESIS - Leibniz Institute for the Social Sciences	16	0.00148341
32	Equinox	Equinox Publishing	16	0.00386484
33	Copernicus	Copernicus Publications	15	0.00155203
34	LRP	Lynne Rienner Publishers	15	0.00152791
35	HUP	Helsinki University Press	15	0.00137502
36	Nomos	Nomos Verlagsgesellschaft mbH & Co. KG	13	0.00143717
37	AAAS	American Association for the Advancement of Science	13	0.00156797
38	AAA&S	American Academy of Arts & Sciences	12	0.00136953
39	AAG	Association of American Geographers	12	0.0014002
40	PA-UBC	Pacific Affairs, University of British Columbia	11	0.00149935

Table C1. Journal Rankings and Impact Values.

†*Note: Only the top 40 journals are shown.*

The information presented in this table not only highlights the diversity of publishers contributing to the topic but also provides a quantitative assessment of their influence. This analysis enables a better

understanding of how the academic publishing landscape shapes the evolution of international security studies and supports informed decision-making in research dissemination and policy development.

Statements and Declarations

Funding

The University of Warsaw financed the research.

Acknowledgements

I would like to thank Alexander Yakimenko for the valuable suggestions and insightful discussions that greatly contributed to shaping the direction and depth of this research.

Footnotes

[1] For more information, see https://en.wikipedia.org/wiki/Cosmic_inflation.

[2] In cosmology, the e-folding time scale is the proper time during which the length of a patch of space or spacetime increases by a factor of e.

[3] https://en.wikipedia.org/wiki/Cosmic_inflation

[4] https://en.wikipedia.org/wiki/Curvaton

References

1. ^*Griffiths TL, Steyvers M. (2004). "Finding scientific topics". Proceedings of the National Academy of Sciences. 101 (suppl_1): 5228–5235. doi:10.1073/pnas.0307752101.*
2. ^*Wu S, Junior B. (2023). "Emerging technologies and global health: A systematic review generating bibliometric evidence for innovation management". BMJ Innovations. 9: bmjinnov–2022. doi:10.1136/bmjinnov-2022-001064.*
3. ^*Chen C. (2017). "Science mapping: A systematic review of the literature". Journal of Data and Information Science. doi:10.1515/jdis-2017-0006.*
4. ^*Cobo MJ, López-Herrera AG, Liu X, Herrera F. (2011). "Science mapping software tools: Review, analysis, and cooperative study among tools". Journal of the American Society for Information Science and Tec*

hnology. doi:10.1002/asi.21525.

5. a, b, c *Andrea Martini, Alessio Cardillo, Paolo De Los Rios. (2018). Entropic selection of concepts unveils h idden topics in documents corpora. ArXiv. Available from: https://arxiv.org/abs/1705.06510.*
6. a, b *Chumachenko A, Kreminskyi B, Mosenkis I, Yakimenko A. (2022). "Dynamical entropic analysis of sc ientific concepts". Journal of Information Science. 48 (4): 561–569. doi:10.1177/0165551520972034.*
7. ^ *Hubert Wagner, Paweł Dłotko, Marian Mrozek. (2012). Computational topology in text mining. In: Co mputational topology in image context: 4th international workshop, CTIC 2012, bertinoro, italy, may 28 -30, 2012 proceedings.: Springer pp. 68–78.*
8. ^ *Hubert Wagner, Paweł Dłotko. (2014). Towards topological analysis of high-dimensional feature spac es. Computer Vision and Image Understanding. 121: 21–26.*
9. ^ *Shafie Gholizadeh, Armin Seyeditabari, Wlodek Zadrozny. (2018). Topological signature of 19th centur y novelists: Persistent homology in text mining. Big Data and Cognitive Computing. 2 (4): 33.*
10. ^ *Salnikov V, Cassese D, Lambiotte R, Jones NS. (2018). "Co-occurrence simplicial complexes in mathem atics: Identifying the holes of knowledge". Applied Network Science. 3: 1–23.*
11. a, b, c *Meila M. (2007). "Comparing clusterings – an information based distance". Journal of Multivariat e Analysis. 98: 873–895. doi:10.1016/j.jmva.2006.11.013.*
12. ^ *Herbert Edelsbrunner, Hubert Wagner. (2016). Topological data analysis with bregman divergences. ar Xiv preprint arXiv:160706274.*
13. ^ *Kvalseth T. (2017). "On normalized mutual information: Measure derivations and properties". Entrop y. 19: 631. doi:10.3390/e19110631.*
14. ^ *Yasuichi Horibe. (1985). Entropy and correlation. IEEE Transactions on Systems, Man, and Cybernetics. SMC-15: 641–642. Available from: https://api.semanticscholar.org/CorpusID:22776467.*
15. ^ *Alexander Kraskov, Harald Stögbauer, Ralph G. Andrzejak, Peter Grassberger. Hierarchical clustering b ased on mutual information. 2003. Available from: https://arxiv.org/abs/q-bio/0311039.*
16. ^ *Havel TF. (1991). "Some examples of the use of distances as coordinates for euclidean geometry". Journ al of Symbolic Computation. 11 (5): 579–593. doi:10.1016/S0747-7171(08)80120-4.*
17. ^ *Astafiev A, Prokofyev R, Guéret C, Boyarsky A, Ruchayskiy O. (2012). "ScienceWISE: A web-based inter active semantic platform for paper annotation and ontology editing". In pp. 392–396. doi:10.1007/978- 3-662-46641-4_33.*
18. ^ *Roman Prokofyev, Gianluca Demartini, Alexey Boyarsky, Oleg Ruchayskiy, Philippe Cudré-Mauroux. O ntology-Based Word Sense Disambiguation for Scientific Literature. In: David Hutchison, Takeo Kanad*

e, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, et al.editors. Advances in Information Retrieval.: Berlin, Heidelberg: Springer Berlin Heidelberg 2013. pp. 594–605. doi:10.1007/978-3-642-36973-5_50. ISBN 978-3-642-36972-8 978-3-642-36973-5

19. ^*Palchykov V, Gemmetto V, Boyarsky A, Garlaschelli D. (2016). "Ground truth? Concept-based communities versus the external classification of physics manuscripts". EPJ Data Science. 5 (1): 28. doi:10.1140/epjds/s13688-016-0090-4.*
20. ^*Andrea Martini, Artem Lutov, Valerio Gemmetto, Andrii Magalich, Alessio Cardillo, et al. ScienceWISE: Topic Modeling over Scientific Literature Networks. arXiv 2016.*
21. ^*Edelsbrunner H, Ölsböck K, Wagner H. (2024). "Understanding higher-order interactions in information space". Entropy. 26 (8). doi:10.3390/e26080637.*
22. ^*Newman M. (2004). "Power laws, pareto distributions and zipf's law". Contemporary Physics. 46. doi:10.1080/00107510500052444.*

Declarations

Funding: The University of Warsaw financed the research.

Potential competing interests: No potential competing interests to declare.