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 introduces a novel methodology for analyzing the evolution of scientific topics through the lens of information geometry. Using mutual entropy-based distance metrics, the approach captures dynamic relationships between scientific concepts over time, offering insights beyond traditional keyword-based analyses. The proposed framework quantifies the influence of publications, institutions, and countries on topic dynamics using normalized velocity matrices and geometric compression measures of knowledge networks. Applying the methodology to data sets from ArXiv and JSTOR, we identify patterns in topic evolution, agent impact, and interdisciplinary influences, emphasizing the utility of entropy-based information-theoretical metrics in understanding the complex dynamics of scientific discourse. The findings highlight applications in strategic planning for academic journals, funding agencies, and research institutions, enabling data-driven decision-making to foster emerging research trends and interdisciplinary collaborations.

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 policymakers 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. While traditional methods such as bibliometric analysis have been instrumental

 doi.org/10.32388/O5NMBG

in tracking trends, they often fall short of capturing 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-theoretical metrics, researchers can explore the geometry and topology of knowledge networks to reveal latent patterns, structural deficiencies, and knowledge gaps. These methods, such as persistent homology and simplicial complexes, have been successfully applied to co-occurrence networks, feature spaces, and text corpora, demonstrating their potential to uncover hidden relationships and bridge knowledge gaps[7][8][9][10]. The study of information topology not only enhances our understanding of knowledge production but also potentially provides actionable insights into the diffusion and evolution of information across disciplines.

This study introduces a novel methodology to analyze the dynamics of scientific topics within knowledge networks. Specifically, we utilize the variation of information (VI) distance metric[11], a measure grounded in information theory, to quantify multiscale geometric-topological properties and higher-order interactions. Unlike non-Euclidean metrics, such as those derived from Bregman divergences, which require customized tools[12], the Euclidean nature of the VI metric allows us to apply well-established computational frameworks, simplifying its integration into existing pipelines. Our approach enables the exploration of topic volume compression as a form of simplicial complex volume dynamics and facilitates the identification and analysis of knowledge brokers—agents that bridge gaps within knowledge networks.

By applying this methodology to datasets from ArXiv and JSTOR, we aim to demonstrate its utility in quantifying the effects of bridging knowledge gaps, assessing interdisciplinary influences, and identifying the impact of various agents, such as institutions and journals. 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 underpinnings, methodological implementation, and empirical findings of this study. By addressing both the geometry and topology 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, $d(X,Y) = 1$. Notably, 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 fixed document corpus. However, comparing distances between concepts at different times when the size of the relevant documents is changing necessitates an additional normalization factor:

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

where N^t is the total number of documents in the corpus published up to time t.

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 velocity matrices, provide additional insight into the temporal evolution of the network.

We compute the velocity matrix by taking the differences between elements of the distance matrices at different time points t_n. Using a constant time interval $\Delta t = t_{n+1} - t_n$, set to one year, reduces noise and computational overhead. For simplicity, we will 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:

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

The result from Eq. (5), the velocity matrix $\tilde{V}$, contains only negative elements, representing the increasing mutual information between concepts as new documents emerge. By computing a Minimum Spanning Tree (MST) of this velocity matrix, we can identify elements and patterns corresponding to the 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 often observed in social science studies and indicative of scale-free network behavior.

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 tree-spanning networks of normalized velocity matrices. This observed relaxation suggests that, without reinforcement from new publications, certain connections may gradually weaken over time due to the increasing total number of documents in the corpus. By applying normalization, we remove the influence of database growth, enabling the analysis to focus solely on network compression driven by entropic forces introduced by new publications.

The effect of excessive knowledge network compression, driven by additional information in emerging publications, can be quantified through the geometric properties of the network. This

approach provides a means of measuring the high-dimensional impact of these sources on the structure of the knowledge network. The metric $\tilde{d}$ enables us to interpret the knowledge network with $n+1$ concepts (nodes) as an n-dimensional simplex with a well-defined volume[16]:

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

To evaluate the impact on the shape of the knowledge network with n concepts from a particular subset of documents, we calculate the normalized relative difference of the topic volume (see Appendix 6):

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

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 *agents* identified through document metadata. The resulting difference is normalized by the number of relevant documents N_r associated with the agent, where each publication contains at least two concepts of the considered n-dimensional topic. When $\Delta \tilde{V}_n / N_r$ approaches zero, the agent's impact is negligible, suggesting that the filtered and unfiltered volumes are nearly identical. In contrast, a higher value of $\Delta \tilde{V}_n / N_r$ reflects a stronger influence of the agent, with the filtered volume $\tilde{V}_n^f$ being significantly larger than the unfiltered volume $\tilde{V}_n^a$, indicating a noticeable compression of the structure of the knowledge network due to the presence of the agent's publications.

In the following section, we first analyze the dynamics of topic network compression by calculating various parameters of normalized and nonnormalized velocity matrices, as well as examining the evolution of network volume as a function of emerging documents. Second, we assess the structure and quantify the impact of agents, such as journals and countries, using the normalized relative volume measure. Examples are drawn from topics in physics and international relations to 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 information was obtained from the ScienceWise platform (now ProphyScience, https://www.prophy.science/)[17]. All JSTOR documents were preprocessed using Google BigQuery to extract keyword metadata sections, and the cleaned collection of keywords was matched with the frequencies of n-grams (uni-, bi- and trigrams) extracted from the full texts of the documents. The metadata for all publications used in our analysis was enriched with information from the Web of Science database by matching DOIs and publication titles.

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 on this 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. 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.

For example, following this algorithm when the parameters are set to $n = 5\%$ and $b = 0.98$, we extract 268 related concepts for the "Reheating" seed concept associated with cosmic inflation in 2018.[1] In Figure 1, we present the semantic landscape of the related topic by illustrating a 2D Delaunay mesh of a corresponding graph[21].

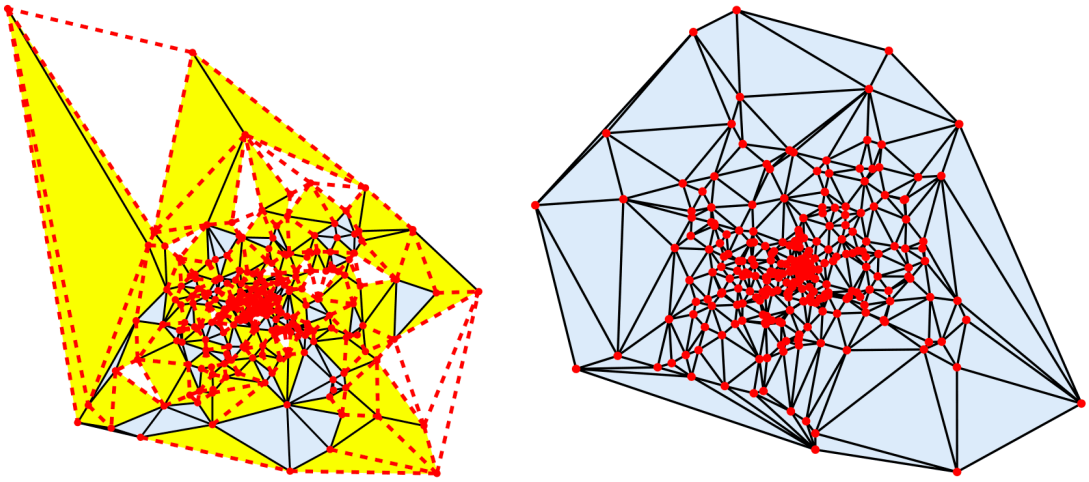

Figure 1. Two-dimensional Delaunay mesh of the normalized "Reheating" topic distance matrix, calculated for 2002 (left) and 2018 (right). Blue triangles represent areas where all three vertices are connected by edges with distances below the specified normalized cut-off distance $\tilde{a} = 0.939042$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

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 components of the velocity matrix, calculated using Eq.(5) for non-normalized distances. In this context, vertex order reflects the *dynamic connectivity* of a concept, indicating whether its mutual information with other concepts in the network is increasing or decreasing. As illustrated in Figure 2 (left), converging subnets tend to have smaller connectivity orders than their relaxation counterparts, highlighting the local character of topic updates. Each concept in the topic simultaneously belongs to both converging and relaxation subnetworks, with its dynamic connectivity order changing in response to signals from new publications. Since Algorithm 1 fixes the number of concepts within a topic, the number of possible links between concepts is also fixed. This constraint results in an almost symmetric distribution of vertex orders in the convergent and divergent components of the velocity matrix at any given time.

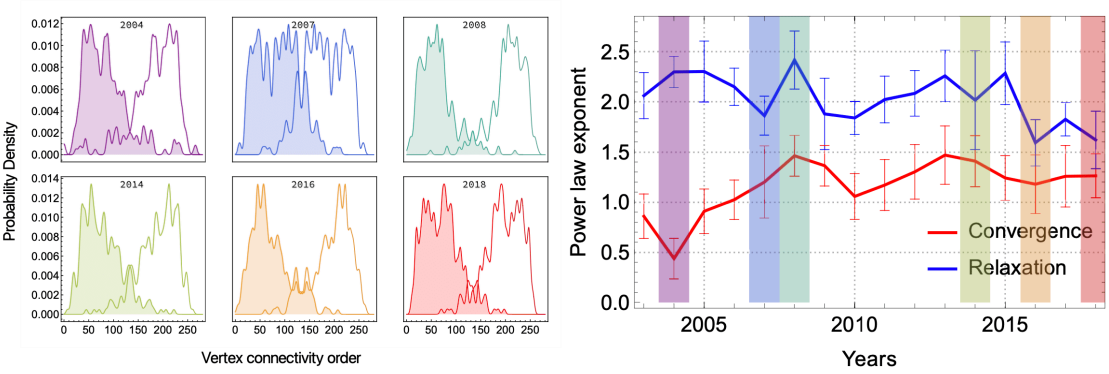

Figure 2. Kernel-smoothed vertex order probability distribution for the positive and negative components of the velocity matrix associated with the evolution of the "Reheating" topic (left). Evolution of the power-law exponent for vertex order distributions in the spanning tree network of the convergent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

The velocity matrix calculated using normalized VI distances contains only negative elements, indicating the continuous convergence of distances within the corresponding knowledge network as the number of emerging documents increases. Each node in the normalized velocity graph has a connectivity order equal to the total number of nodes in the topic, resulting in a uniform distribution of the connectivity order of the vertex in the spanning tree network derived from the normalized VI distances. However, the distribution of relative velocities within the 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 uniform. From Figure 3 (right), we can observe that the size of the topic volume calculated using Eq. (7) is a function of the total number of relevant documents, and this size decreases as new documents emerge.

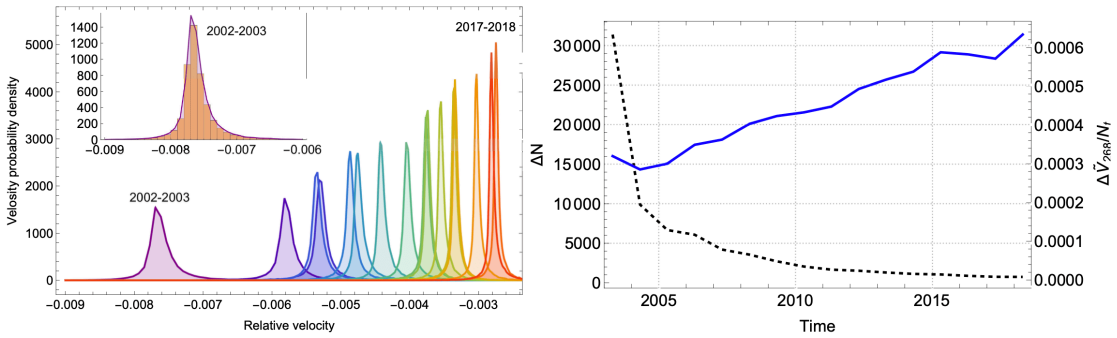

Figure 3. The dynamics of the velocity distribution in the 'Reheating' knowledge network topic (left) and the evolution of the critical exponent for the corresponding spanning tree networks, calculated with a one-year time step. The comparison of changes in the topic's relative simplex volume over time (black dashed line) is shown alongside the total number of related documents (blue line) published annually in ArXiv.

This 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, introducing new concepts and connections leads to significant adjustments in the network geometry. The most relevant connections are established as the topic matures, 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.

Using the topic volume as the volume of a multidimensional 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. The impact structure can also be retrieved from the analysis of the generalized 'velocity' matrix-like approach when we trace not timely changes but the effect that a specific group of documents makes on the structure of the studied topic. Document metadata allows for broad classification, and our approach allows us to quantify the impact of some generalized class of documents that we associate with some abstract 'agent'. It can be an author, journal, publisher, grant agency, institution, country, or geographical region. We can identify the impact of a specific topic on another topic, as long as we can identify relevant groups of documents in a particular database.

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. 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 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.

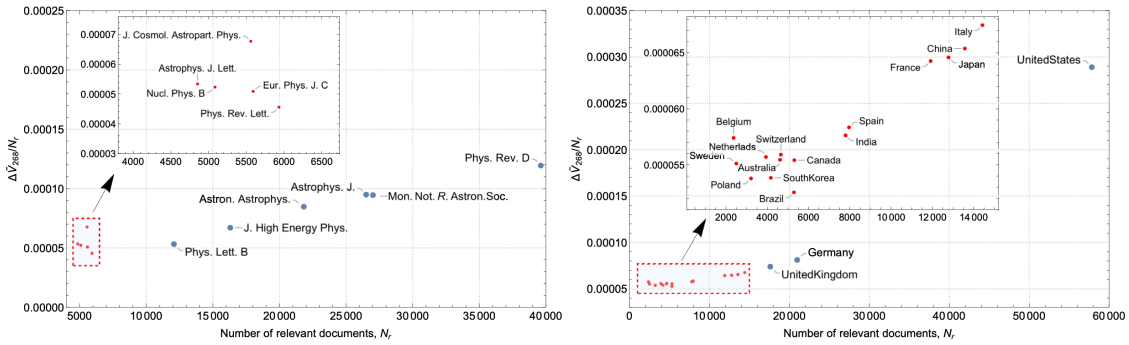

Figure 4. The impact on the 'Reheating' topic knowledge network with $n = 268$ concepts from journals (left) and countries (right), measured as the relative change in the topic's simplex volume based on the number of relevant documents published in ArXiv from 1990–2018.

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

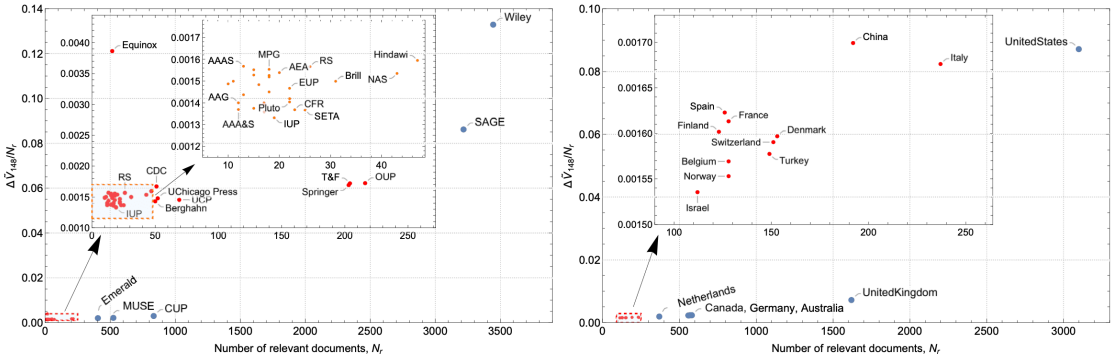

Figure 5. The value of the impact on the "International security" topic knowledge network with $n = 148$ concepts from documents related to countries (right) and publishers (left), measured as relative topic simplex volume change per document published in the period 2010-2024.

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.

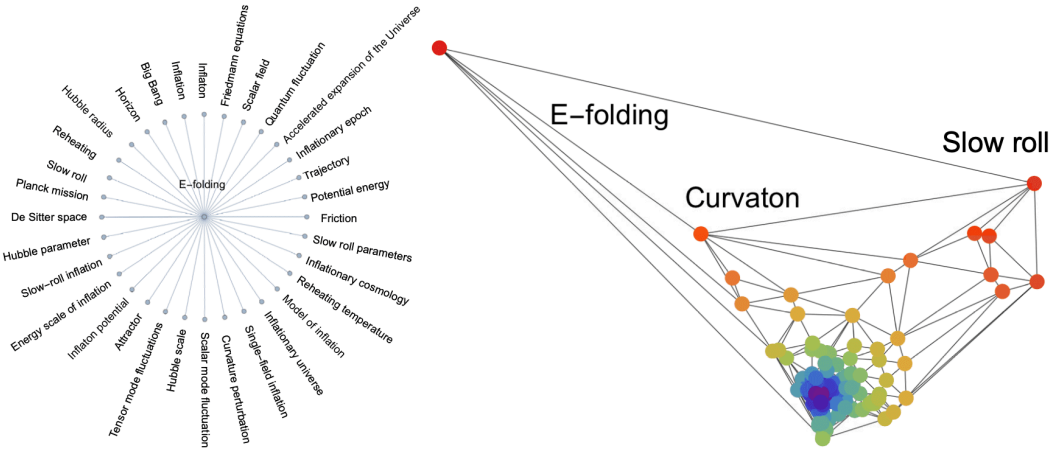

Figure 6. The structure of the largest impact made on the "Reheating" topic network structure from documents whose corresponding author has an affiliation in the United States of America. The node of the highest connectivity and its related vertices are extracted from the graph difference spanning tree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

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.

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). 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. Notably, the triangle formed by the concepts "E-folding," "Slow roll," and "Curvaton" represents the most discussed terms within the "United States" document category.

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 agents—such as countries, journals, and institutions—on the geometry of these networks.

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 agents, the methodology quantifies the influence of authors, institutions, journals, and nations on shaping topic dynamics. This capability highlights the role 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 in ArXiv to international relations topics in JSTOR, 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.

Looking ahead, future research should address the universality of these findings in diverse datasets and disciplines. Incorporating additional normalization techniques 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.

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 uses JSTOR and ArXiv (http://arxiv.org), which are publicly available sources of information containing metadata and texts of scientific documents. In JSTOR (https://constellate.org/), the texts of the documents are parsed, and uni-, bi-, and trigrams are extracted, making it possible later to match available dictionaries and collect the ontologies term frequencies. The scientific dictionary provided by the ScienceWise platform for scientific analysis is not openly available due to the terms of the conducted agreement with this 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.

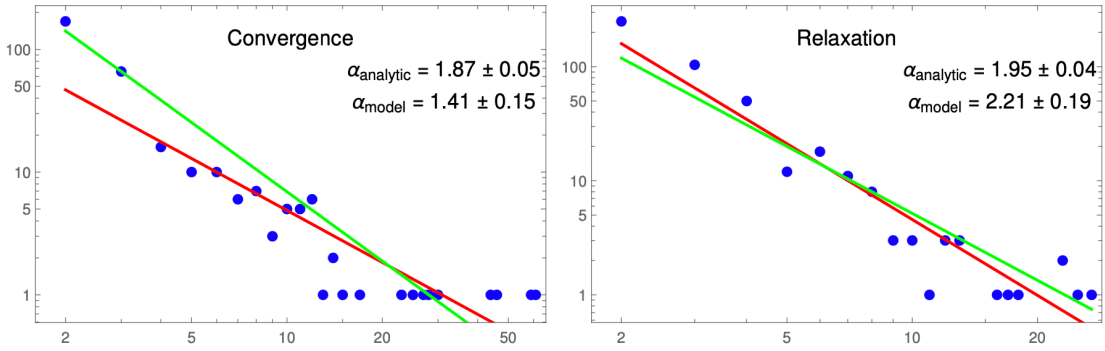

Figure A1. Log-log plot for vertex orders within the "Reheating" related spanning tree network of the positive and negative parts of the velocity matrix for 2017-2018. The x-axis represents the vertex order (degree), and the y-axis shows the number of nodes. The vertex order distribution follows a power-law $k^{-\alpha}$. 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.). The volumes $\tilde{V}_a$ and $\tilde{V}_f$ are calculated based on the 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 the 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 N_a or N_f in 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).

The 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 associated knowledge production. By capturing these effects, it offers a generalized perspective on estimating how scientific publications shape the structure of the knowledge network.

Footnotes

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

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 Technology. doi:10.1002/asi.21525.*
5. ^ *Andrea Martini, Alessio Cardillo, Paolo De Los Rios. (2018). Entropic selection of concepts unveils hidden topics in documents corpora. ArXiv. Available from: https://arxiv.org/abs/1705.06510.*
6. ^ *Chumachenko A, Kreminskyi B, Mosenkis I, Yakimenko A. (2022). "Dynamical entropic analysis of scientific 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: Computational 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 spaces. Computer Vision and Image Understanding. 121: 21–26.*
9. ^ *Shafie Gholizadeh, Armin Seyeditabari, Wlodek Zadrozny. (2018). Topological signature of 19th century 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 mathematics: Identifying the holes of knowledge. Applied Network Science. 3: 1–23.*
11. a, b, c *Meilă M. (2007). "Comparing clusterings – an information based distance". Journal of Multivariate Analysis. 98: 873–895. doi:10.1016/j.jmva.2006.11.013.*
12. ^ *Herbert Edelsbrunner, Hubert Wagner. (2016). Topological data analysis with bregman divergences. arXiv preprint arXiv:160706274.*
13. ^ *Kvalseth T. (2017). "On normalized mutual information: Measure derivations and properties". Entropy. 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". Jour nal of Symbolic Computation. 11 (5-6): 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 interac tive 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_5 0. 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 communi ties versus the external classification of physics manuscripts". EPJ Data Science. 5 (1): 28. doi:10.1140/epj ds/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 informati on 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.