SCIENTIFIC COMMUNITY

Neural embeddings of scholarly periodicals reveal complex disciplinary organizations

Hao Peng[1], Qing Ke[2], Ceren Budak[1], Daniel M. Romero[1,3,4], Yong-Yeol Ahn[5,6]*†

Understanding the structure of knowledge domains is one of the foundational challenges in the science of science. Here, we propose a neural embedding technique that leverages the information contained in the citation network to obtain continuous vector representations of scientific periodicals. We demonstrate that our periodical embeddings encode nuanced relationships between periodicals and the complex disciplinary and interdisciplinary structure of science, allowing us to make cross-disciplinary analogies between periodicals. Furthermore, we show that the embeddings capture meaningful "axes" that encompass knowledge domains, such as an axis from "soft" to "hard" sciences or from "social" to "biological" sciences, which allow us to quantitatively ground periodicals on a given dimension. By offering novel quantification in the science of science, our framework may, in turn, facilitate the study of how knowledge is created and organized.

INTRODUCTION

Since the formalization of science, scholarly periodicals, such as academic journals and proceedings, have become the primary loci of scientific activities (*1*–*4*). Periodicals are not only the conduits of scientific communication but also distributed repositories of scientific knowledge organized around topical niches and disciplines (*4*, *5*). Therefore, scholarly periodicals have been considered the fundamental units for investigating the structure and evolution of science (*6*–*10*).

Moving beyond manually curated classification systems, previous studies leveraged citation and other metadata to capture relationships between periodicals in the form of similarity matrices or networks, which led to algorithmically created "maps of science" and insights into the structure of disciplinary organization (*6*–*9*, *11*, *12*).

Yet, although the vector-space model (*13*) could provide a powerful framework for quantitative inquiries by allowing algebraic operations among periodicals, it has not been adopted much in the traditional approaches. The vector representation based solely on the explicit connections suffers from sparsity; for instance, using inter-citation or co-citation as a similarity measure produces a sparse matrix where most elements are zeros (*6*). Incorporating indirect relationships would pose many choices for the metrics and require handling of a large, dense similarity matrix.

Recent advancement in machine learning has demonstrated that neural embedding techniques offer a powerful solution to these issues. Neural embedding is a family of techniques for obtaining compact, dense, and continuous vector-space representations of entities that can efficiently encode multifaceted relationships between those entities, and has become a core ingredient in modern machine learning (*14*). The embedding approach, instead of focusing on the explicit relationships between entities, aims to learn concise representations that capture both explicit and implicit relationships between the entities. Although its precursor, the vector-space model, was developed many decades ago (*13*), the neural network approach—because of its flexibility, efficiency, and robustness (*15*)—has recently produced many breakthroughs.

Because it was demonstrated that word embeddings can encode rich semantic relationships between words as geometrical relationships in low-dimensional vector space (*16*–*20*), the embedding models have offered novel opportunities and solutions to challenging problems, including language evolution (*21*, *22*), gender and stereotypes (*23*, *24*), culture and identities (*25*, *26*), and even the prediction of material properties (*27*).

Furthermore, the idea of training vector-space embedding using neural networks is not limited to words—it has been adopted to other entity types, including sentences, paragraphs, documents, images, and networks (*28*–*31*).

Here, we propose a network embedding method to learn dense and compact vector-space representations of periodicals from the paper citation network. We show that the periodical embeddings can effectively encode the complex organization of knowledge in science, which allow us to perform novel quantitative analyses such as making cross-disciplinary analogies between periodicals.

Namely, we show that our dense embedding approach can (i) better capture similarity between periodicals than traditional alternatives, (ii) produce a high-resolution map of disciplinary organization that can provide insights into the existing classification systems, particularly regarding interdisciplinary research areas, (iii) allow us to make meaningful analogies between periodicals, and (iv) identify robust spectra of periodicals along conceptual dimensions such as the soft-hard science axis and the social-biological science axis.

Our embedding method builds on the DeepWalk and node2vec model (*30*, *31*), which are a direct adaptation of the word2vec model in the context of networks. In this framework, random walks on the network are considered as "sentences." Instead of using the network of periodicals, our method leverages the richer and higher-order citation network of papers to learn the representations of periodicals (see Materials and Methods).

Let us sketch the key idea. Imagine reading a paper from a field that you are unfamiliar with. To understand this paper, you may need to read another paper from the reference list, which, in turn, may prompt you to read another earlier paper, taking you down a

[1]School of Information, University of Michigan, Ann Arbor, MI 48109, USA. [2]Center for Complex Network Research, Northeastern University, Boston, MA 02115, USA. [3]Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Mi 48109, USA. [4]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, NI 48109, USA. [5]Center for Complex Networks and Systems, Luddy School of Informatics, Computing, Bloomington, IN 47408, USA. [6]Indiana University Network Science Institute, Bloomigton, In 47408, USA.
*Corresponding author. Email: yyahn@iu.edu
†Present address: Connection Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

rabbit hole-like citation trail. We hypothesize that these citation trails, created from references between papers, capture natural sequences in the citation network. Now, by looking at the periodicals where each of the papers in the citation trail was published, we can obtain a trail of periodicals. Here, we consider each periodical as a "word" and each trail as a "sentence." If we apply the word2vec model to these sentences, it lets us learn embeddings that encode the semantic relationships among periodicals. Similar to the case of word embeddings, periodicals with similar context in the citation trails would have similar vector-space representations. Note that, instead of using random walks on the citation network of periodicals, we leverage richer and higher-order trajectories from the lower-level paper citations to enrich the output embeddings [see information gained from higher-order trajectories (*32*)].

RESULTS

We applied our method to a citation network of 53 million papers and 402 million citation pairs built from the Microsoft Academic Graph (MAG) (see Materials and Methods).

As a result, we obtained a 100-*d* unit vector for each of the 20,835 periodicals. Our embeddings offer natural ways, i.e., the cosine similarity between vectors, to measure similarities between periodicals. For instance, the two closest periodicals to *Proceedings of the National Academy of Sciences* are *Nature* and *Science*, and the two closest periodicals to *American Sociological Review* are *Social Forces* and *American Journal of Sociology* (see fig. S3 for the top list and other examples).

Validating the embedding space

We compare our dense periodical embeddings ("p2v") with two citation-based vector-space models. Specifically, we construct an adjacency matrix representing the citation counts between 24,020 periodicals. The first baseline is a citation vector ("cv") model, for which we assign a 48,020-*d* vector to each periodical by concatenating its in-degree vector and its out-degree vector (both are normalized to the unit length). In contrast to our method that compresses the information contained in the paper citation network into low-dimensional dense periodical embeddings, this citation vector model makes use of the citation network itself and each periodical is represented by its citation pattern with respect to every other periodical. The second baseline is a periodical similarity matrix obtained by applying the Jaccard similarity measure to the periodical citation matrix ("jac"), which is the best model reported in (*6*). In this similarity matrix (also a sparse matrix), an entry m_{ij} represents the total number of citations between periodicals i and j, normalized by their total number of citations to other periodicals. Each periodical is then represented as its row vector.

We evaluate our embeddings against the two vector-space models and other baseline methods in three tasks: (i) capturing the similarities between pairs of journals in the same discipline, (ii) comparing the ranking of similar journals to that perceived by experts, and (iii) predicting the discipline category for journals. We focus on 12,780 journals for which we have the discipline information through the matching with the University of California San Diego (UCSD) map of science catalog (table S1).

The first task examines how well the embeddings can systematically capture journal similarities across disciplines. We randomly sample 100,000 journal pairs for four groups: (i) random pairs, (ii) pairs in different disciplines, (iii) pairs in the same discipline, and (iv) pairs in the same subdiscipline. Figure 1 (A to C) shows the distribution of cosine similarities for journal pairs calculated based on the three vector-space models. According to the citation vector method and the Jaccard similarity matrix, most journal pairs in the same discipline (or even in the same subdiscipline) have a similarity score of 0 or close to 0 (the lowest possible value that can be produced by the two methods because of nonnegative vector elements), highlighting the primary weakness of the sparse encoding approach: It fails to capture meaningful similarity variation across many pairs. By contrast, our embeddings provide a wide range of similarity scores (from −0.5 to 1) for random journal pairs (including pairs in different disciplines).

The mean similarity values for four groups of journal pairs are 0.02, 0.03, 0.10, and 0.28 based on the citation vector model; 0.008, 0.009, 0.046, and 0.175 based on the Jaccard similarity matrix; and 0.07, 0.03, 0.25, and 0.54 based on our embedding (the corresponding mean values are statistically different from each other; P values are negligibly small). We also compute Kullback-Leibler (KL) divergence by estimating the probability density function of similarity scores using the kernel density estimation (with the exponential

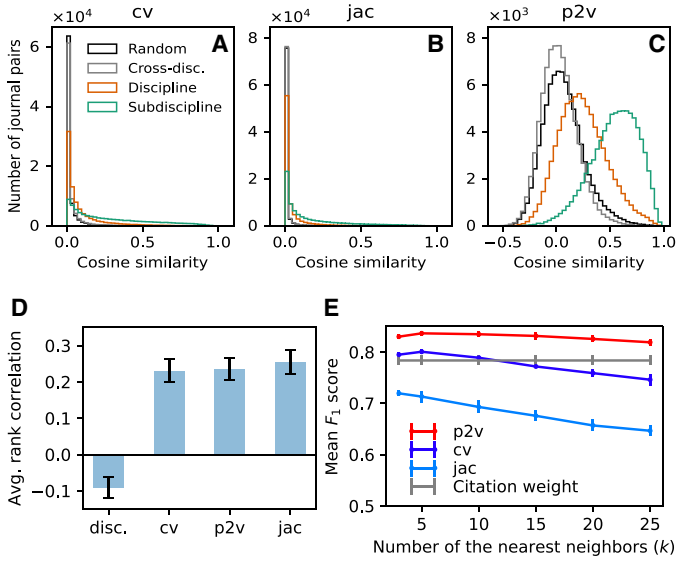

Fig. 1. Model validation. (**A** to **C**), The distribution of cosine similarities for four groups of 100,000 journal pairs calculated based on the citation vector (cv) model, the Jaccard similarity matrix (jac), and our dense periodical embeddings (p2v). The four labels—random, cross-disc., discipline, and subdiscipline—represent random pairs, cross-discipline pairs, within-discipline pairs, and within-subdiscipline pairs. The two sparse embeddings (cv and jac) put most pairs at 0 and thus are not as informative as our dense embedding, which better captures journal similarities and their differences. Compared to random pairs, both the means and the distributions of the other three groups shift more dramatically based on p2v than that based on either cv or jac. (**D**) Average rank correlation coefficient between algorithms and experts in ranking topically similar journals. Target journals with an average pairwise expert agreement above 0.2 are used in the evaluation. The label disc. represents the method that ranks journals in the same discipline based on their PageRank scores. (**E**) F_1 score of the classification task in predicting the discipline category for 12,751 journals (excluding 29 interdisciplinary journals) using the three vector-space models. The results are based on a five-fold cross validation. The "label citation" weight represents the method that predicts the discipline of a journal to be that of its most cited neighbor in the undirected journal citation network. Error bars indicate 95% confidence intervals.

kernel for cv and jac and the Gaussian kernel for p2v). The distribution of within-discipline pairs shifts more dramatically from that of random pairs based on our embeddings (the KL divergence is 0.34 for p2v versus 0.25 for cv and 0.11 for jac). The displacement for within-subdiscipline pairs is even larger (KL divergence: 2.06 versus 1.57 or 1.07), which is also true for cross-discipline pairs (KL divergence: 0.026 versus 0.002 or 0.000). This result underlines the benefits of dense and continuous embedding over sparse encoding in capturing journal similarities across disciplines defined in an existing journal classification system.

Our second task is ranking periodicals based on their topical similarity to a given target periodical. In addition to our method (p2v) and the two baseline vector-space models (cv and jac), we use another baseline (noted as "disc.") that ranks periodicals in the same discipline based on their PageRank scores on the full directed and weighted periodical citation network. The rankings produced by these models are compared to a reference ranking that was constructed from an expert survey.

The survey, which was distributed over the authors' institutions (see Materials and Methods), asks experts to rank a set of candidate periodicals for a given target periodical based on their topical similarity. We then compare the algorithms' rankings with that given by experts. Figure 1C shows the average Kendall's rank correlation coefficient between each algorithm and experts. The three vector-space models perform similarly better than the first baseline. Although their performances are similar, our embeddings are orders of magnitude more computationally efficient than the citation vector model and the Jaccard similarity matrix in terms of time and space complexity due to low dimensionality (100 versus 48,040 or 24,020).

The low correlation between algorithms and experts is mainly because experts themselves have high disagreement—the average pairwise rank correlation per target journal is 0.14. This may not be unexpected given the subjective nature of the task (see fig. S5 for an example), which is also evidenced by the fact that, on average, 41.5% of candidate journals were placed into the "Unfamiliar Journals" bucket by experts.

Last, our third task tests the predictability of discipline category of a given periodical based on its neighbors. We focus on 12,751 journals (excluding 29 interdisciplinary ones). We compare our embedding to the same citation vector model (cv), the Jaccard similarity matrix (jac), and another baseline (labeled as "citation weight"), which predicts the discipline of a target journal to be that of its highest-strength neighbor in the undirected journal citation network, where the edge weights are defined as the total number of citations between two journals (the undirected version performs better than the two directed versions). For p2v, cv, and jac, we use the k-nearest neighbors algorithm based on vector similarities for the prediction task. Figure 1D shows that our embeddings can more accurately predict journals' discipline category. In other words, the neighbors in the embedding space tend to belong to the same discipline and this tendency is stronger in our model.

Together, these results indicate that our periodical embeddings, while being much more efficient, can better capture the relationships between journals than the sparse vector-space models based on citations and other baseline approaches.

Disciplinary structure revealed by the periodical embedding

The embeddings of scholarly periodicals also encode the complex disciplinary structure in the knowledge space. Figure 2A presents a two-dimensional (2D) representation of the embeddings of 12,780 journals, providing an overview of the global structure of 13 major scientific disciplines (an interactive version is available at: https://haoopeng.github.io/journals). Although our approach produces continuous—not categorical—representations of periodicals, to facilitate a comparison with a traditional journal classification system, we color each journal in Fig. 2A based on its discipline category designated in the UCSD map of science catalog (*9*). The 13 disciplines defined in the UCSD map still show up as conspicuous regions in our projection. However, it also exposes the nuanced structure and the limitations of the classification approach. For instance, it uncovers interdisciplinary microclusters, such as parasite research or neuroimaging, that cannot be properly captured in the disjoint categories (see Fig. 2, B to D, and figs. S14 to S26 for other examples).

If our embedding is indeed capable of capturing interdisciplinary periodicals, it is reasonable to hypothesize that stronger disagreement about a periodical with a traditional classification indicates stronger interdisciplinarity or wrong/ambiguous classification.

To test this hypothesis, we compare our vector-space map to the UCSD classification system (with 13 major categories) systematically and quantitatively. We apply the k-means algorithm to our embedding vectors and cluster journals into 13 groups (29 multidisciplinary journals were excluded from 12,780 matched journals). These organically discovered clusters are then compared to the 13 major categories in the UCSD classification system by using the element-centric similarity measure (*33*). This method allows us to quantify similarity between two clusterings at the level of individual element, thereby enabling us to quantify disagreement for each periodical.

Figure 2E shows the map of agreement between the clustering based on our embeddings and the UCSD categorizations for 12,751 journals. Those interdisciplinary areas that we highlighted in Fig. 2 (B to D) exhibit strong disagreement.

The distribution of agreement scores for journals in each discipline is multimodal (fig. S7). In other words, although the two clusterings are fairly similar for a large fraction of journals, there are still many whose membership across the two clusterings are distinct, possibly indicating their interdisciplinary nature. We performed a manual evaluation to estimate how clearly a periodical belongs to the discipline defined by the UCSD catalog for both high-agreement and low-agreement journals in three disciplines (see Materials and Methods).

Figure 2F shows that journals with a high degree of agreement between the two clusterings can also be clearly identified in their designated discipline. On the contrary, for about 40% journals on which the two clusterings strongly disagree, their discipline designation in the UCSD catalog is disputable. A manual inspection reveals that many low-agreement journals are interdisciplinary and difficult to be classified into a single category (e.g., *Biostatistics*, *Aggressive Behavior*, and *Cell Biology Education* are classified as "Social Sciences" in the UCSD map).

These results suggest that our periodical embeddings, while agreeing with the UCSD categorization on clearly disciplinary journals (Fig. 2F, top), can identify interdisciplinary journals that are difficult to categorize into disjoint disciplines (Fig. 2F, bottom).

This result shows that the dense periodical embedding is a promising data-driven approach to quantitatively operationalize interdisciplinarity using vector similarity. For instance, one may quantify a paper's degree of interdisciplinarity as the average cosine distance between its cited periodicals.

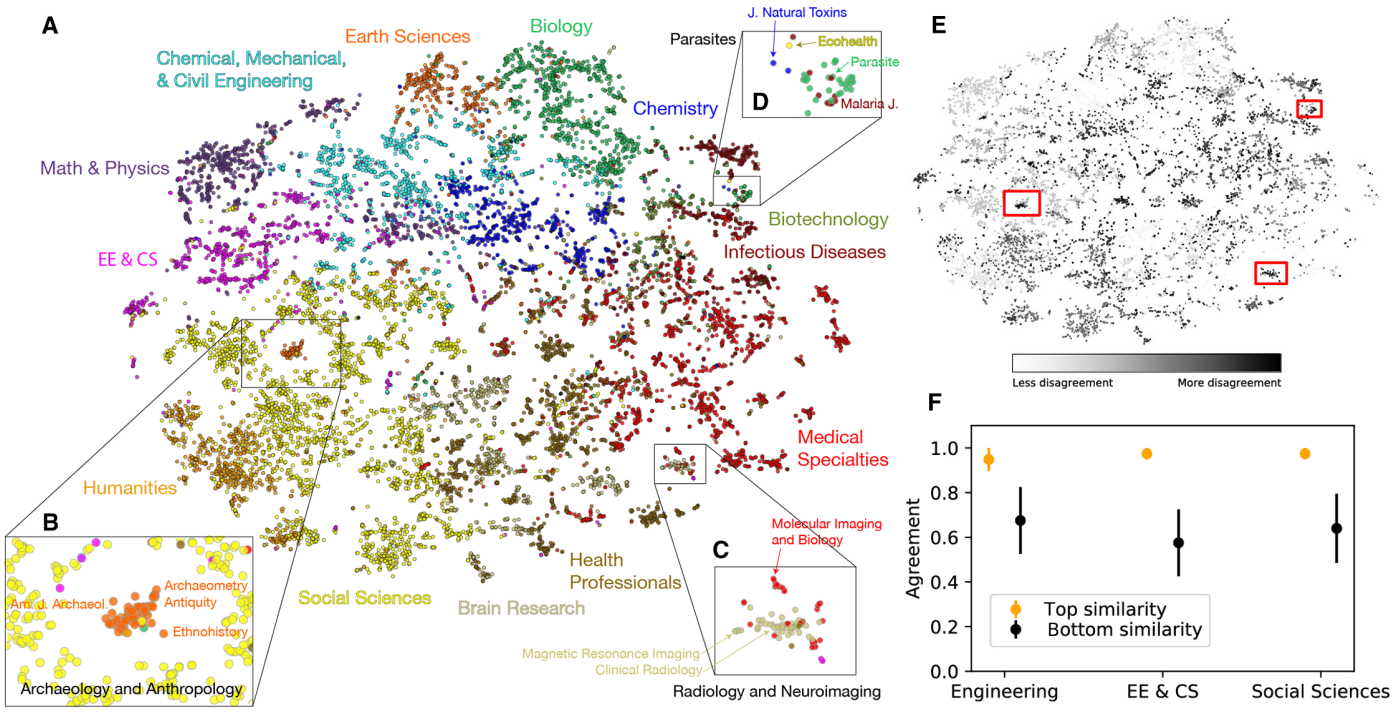

Fig. 2. Periodical embeddings reveal complex disciplinary organizations. (**A**) The two-dimensional (2D) projection of 12,780 journals obtained using *t*-distributed stochastic neighbor embedding (*t*-SNE) (*52*). Each dot represents a journal, and its color denotes its discipline designated in the UCSD map (29 multidisciplinary journals are colored in black). (**B**) Archaeology and anthropology journals, classified as "Earth Sciences," form a distinct cluster with its center closer to "Social Sciences" than the major "Earth Sciences" cluster (verified by cosine distances). (**C**) Group of medical imaging journals comes from "Brain Research," "Medical Specialties," and "EE & CS," highlighting the key role of computer science and engineering in the study of brain imaging. (**D**) Set of parasite-focused journals spans many disciplines, including "Social Sciences" (*Ecohealth*), "Biology" (*Parasites*), "Infectious Diseases" (*Malaria Journal*), and "Chemistry" (*Journal of Natural Toxins*), revealing the multifaceted, highly interdisciplinary nature of parasite research. (**E**) The same map but with a grayscale representing the level of disagreement between the clustering in our embedding space and the discipline categories in the UCSD map. Red rectangles highlight the locations in (B) to (D). (**F**) Agreement between UCSD classifications and our survey. The top (bottom) represents journals with high (low) similarity between the UCSD catalog and a clustering based on our periodical embeddings.

Cross-disciplinary analogies between scholarly periodicals

One of the primary reasons behind the wide adoption of the word2vec model is its uncanny ability to capture semantic relationships geometrically in vector space (*20*, *24*, *26*). The most famous example goes like this: $\mathbf{v}(\text{king}) - \mathbf{v}(\text{man}) + \mathbf{v}(\text{woman}) \approx \mathbf{v}(\text{queen})$. That is, the difference between man and woman (or king and queen) vectors captures the axis of gender, which can be generalized to other gendered nouns such as brother and sister [i.e., $\mathbf{v}(\text{brother}) - \mathbf{v}(\text{man}) + \mathbf{v}(\text{woman}) \approx \mathbf{v}(\text{sister})$] (*16*, *17*).

Can we make similar analogies between scholarly periodicals using our embeddings? For instance, given a periodical pair (A, B), where A is a quintessential Computer Science periodical and B is the one for Sociology, can $[\mathbf{v}(A) - \mathbf{v}(B)]$ capture the axis that runs between Computer Science and Sociology? If that is the case, given a "seed" periodical, we can also use the vector analogy to explore other periodicals that are closer to Computer Science and farther away from Sociology than the seed, or vice versa, using the vector $[\mathbf{v}(B) - \mathbf{v}(A)]$.

We would like to note that one needs to be cautious about the interpretation of word analogies. Commonly, word analogy does not allow duplicates (i.e., all words in the analogy need to be different), which can be misleading in some contexts such as the study of biases in word embeddings (*34*). Here, we maintain this constraint because we specifically aim to discover a new periodical using the analogy.

To demonstrate the possibility of making these cross-disciplinary periodical analogies, we create "analogy graphs," which are constructed by repeatedly performing the vector analogy and taking the best candidate periodical at each step. We first choose two canonical disciplinary periodicals and consider them as the "poles" of an axis going from one discipline to the other. Using the two poles, given a seed periodical, we then iteratively make analogies to the seed and subsequently discovered periodicals.

All identified periodicals, including the seed, can be visualized as a directed network with nodes representing periodicals and links representing the analogical relationships.

Figure 3A shows the analogy graph for *ICWSM* (*The International AAAI Conference on Web and Social Media*) and *KDD* (*ACM SIGKDD Conference on Knowledge Discovery and Data Mining*), produced by applying *JMLR* (*Journal of Machine Learning Research*) and *ASR* (*American Sociological Review*)—two poles of an axis that goes from Sociology to Machine Learning—to each seed (*ICWSM* or *KDD*). Figure 3A reveals a spectrum of periodicals that sit between Sociology and Machine Learning, from a disciplinary sociology journal (*Social Forces*) to interdisciplinary computational social science conferences [e.g., *EMNLP* (*Empirical Methods in Natural Language Processing*) and *IEEE International Conference on Data Mining*], to more method-oriented machine learning conferences [e.g., *ICML* (*The International Conference on Machine Learning*) and *NeurIPS*

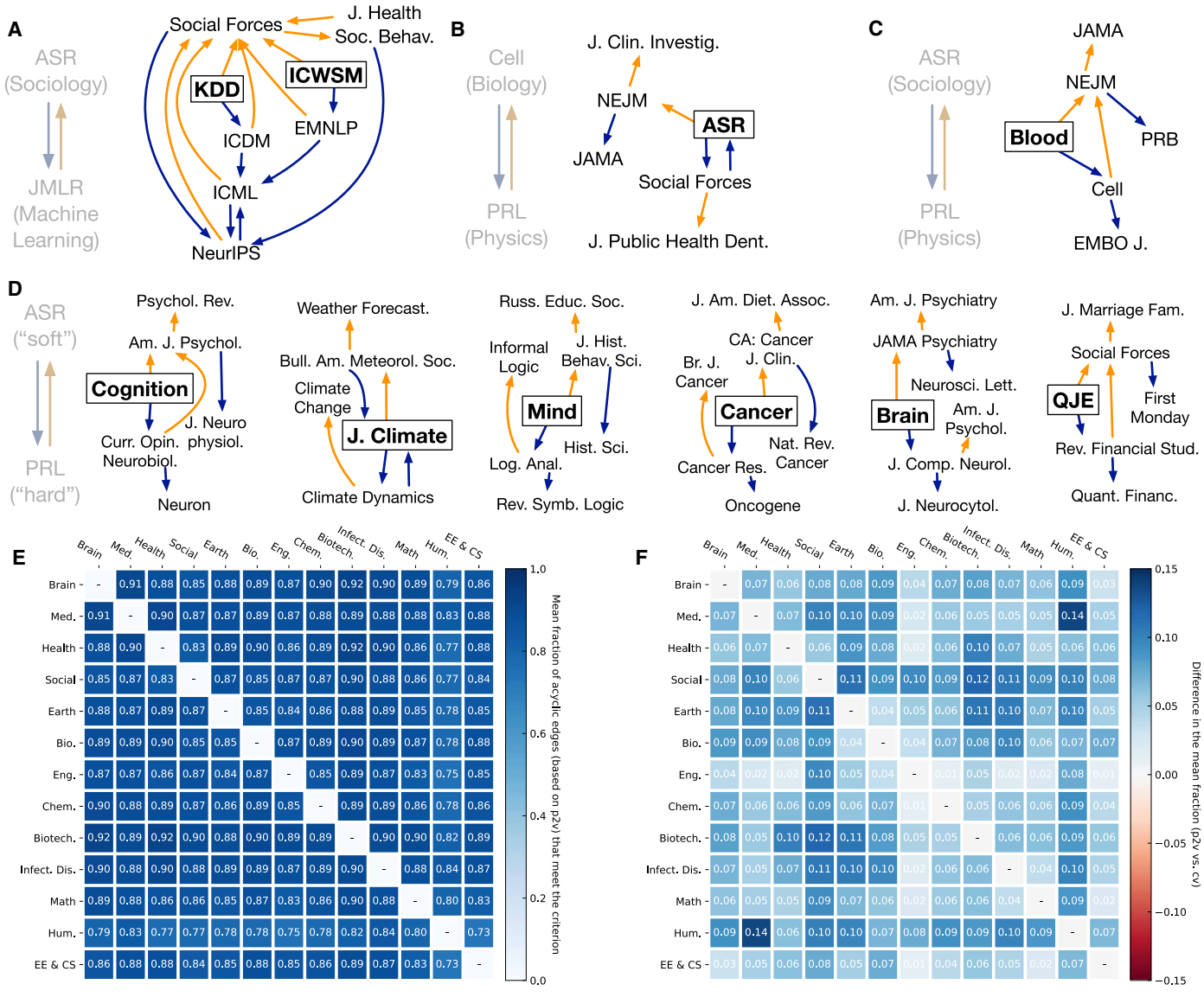

Fig. 3. Analogy graphs between periodicals. (**A**) We apply two poles (*ASR*, *JMLR*) to *KDD* (or *ICWSM*) iteratively to find the most similar periodical at each step via the vector analogy: $\mathbf{v}(X) - \mathbf{v}(ASR) + \mathbf{v}(JMLR) \approx \mathbf{v}(?)$ (blue edges) or $\mathbf{v}(X) - \mathbf{v}(JMLR) + \mathbf{v}(ASR) \approx \mathbf{v}(?)$ (orange edges). Each node has two outgoing edges (blue or orange) representing the two opposite analogies. (**B**) We apply (*Cell*, *PRL*) to *ASR* and only expand periodicals that are one step away from *ASR* to make the graph concise. (**C**) Graph obtained by applying (*ASR*, *PRL*) to *Blood*. (**D**) Similar to (C), for seeds in different disciplines, including "Brain Research" (*Cognition*, *Brain*), "Earth Sciences" (*Journal of Climate*), "Humanities"(*Mind*), "Medical Specialties" (*Cancer*), and "Social Sciences" (*Quarterly Journal of Economics*). (**E**) Average fraction of acyclic edges per analogy graph that satisfy the author overlap criterion for all 1800 analogy graphs (produced by our periodical embeddings, p2v) in each of the 78 discipline pairs. (**F**) Same as (E) but for the differences in the mean values from the analogy graphs produced by cv. For all discipline pairs, the difference is positive and statistically significant (at $P < 0.001$).

(*The Conference on Neural Information Processing Systems*)]. Another analogy graph is obtained by applying the periodical pair [*Cell*, *PRL* (*Physical Review Letters*)] that represents the axis from Biology to Physics, to the seed journal *ASR*, which identifies periodicals with biological flavor—*NEJM* (*The New England Journal of Medicine*)—or more physics flavor—*Social Forces* (Fig. 3B). We apply, in Fig. 3 (C and D), the pair (*ASR*, *PRL*) to periodicals across disciplines; for instance, when applied to *Blood*, we can discover a more "physical" journal (*Cell*) and a more "sociological" journal (*NEJM*). Note that in Fig. 3D, we only identify the most similar periodical that is in the same discipline as the seed during each step.

We then more systematically examine the validity of the periodical analogies with an external dataset—author overlap between periodicals. The intuition is that, as we move away from a periodical (say A) and toward another (say B)—if the analogy works as intended—we will arrive at a periodical that is farther away from A but closer to B, in comparison with the original periodical that we started.

Specifically, for a periodical analogy "$A : B \sim C : D$" [an edge $(C \rightarrow D)$ in an analogy graph produced with A and B as two poles; Fig. 3], we verify whether their author overlaps satisfy the following condition: $\frac{O(C,A)}{O(C,B)} > \frac{O(D,A)}{O(D,B)}$, where $O(P_1, P_2)$ is the number of shared authors—those who have published in both periodicals P_1 and P_2.

That is, as one starts from C and ends up at D by moving further away from A and getting closer to B, we expect that the ratio of author overlap for $O(C, A)$ versus $O(C, B)$ should be larger than that for $O(D, A)$ versus $O(D, B)$.

For example, for the analogy "*ASR*: *JMLR* ~ *EMNLP*: *ICML*," the ratio of (*EMNLP*, *ASR*) author overlap to (*EMNLP*, *JMLR*) overlap should be larger than that between the (*ICML*, *ASR*) author overlap and the (*ICML*, *JMLR*) overlap. We can then identify the acyclic edges that satisfy the criterion and obtain their fraction for any analogy graph (generated with a seed and two poles). Using this pipeline, we compare the quality of periodical analogies generated by our embeddings to that produced by the citation-based sparse encoding model defined previously.

We systematically generate, for each pair of disciplines (D_1, D_2), all 1800 analogy graphs by selecting two poles and the seed from the top 10 journals in D_1 and D_2 (based on their PageRank scores; see "disc." in Fig. 1C). For example, for the pair ("Social Sciences", "EE & CS"), we have 10 journals in each field, which give 100 pairs of poles; for each pair of poles, we have 18 journals (10 + 10 − 2) that can be used as the seeds; we can then generate 1800 analogy graphs for ("Social Sciences", "EE & CS"). We calculate the average fraction of acyclic edges that satisfy the author overlap criterion for all 1800 analogy graphs in each discipline pair (see fig. S8 for an example). We then compare the mean fraction for analogy graphs produced by the two vector-space models (p2v versus cv; see Fig. 1). The results shown in Fig. 3 (C and D) indicate that the periodical analogies produced by our embeddings are better aligned with author overlap between periodicals than those produced by the citation-based sparse vector model, for every pair of the 78 possible discipline pairs.

Extracting conceptual dimensions in disciplinary organizations

The power of embeddings to discover analogical relationships between periodicals prompts us to explore more general conceptual dimensions in the knowledge space, because the two disciplinary poles of a scientific "axis" can be defined not only by a periodical pair but also by two sets of periodicals.

We first pick two general disciplinary areas and calculate their centroids by taking the average of all periodical vectors in each area. Given the two centroid vectors, we obtain an axis that runs from one disciplinary area to the other as we did in the previous examples with individual periodicals. Formally, let $\mathcal{S}^+ = \{\mathbf{v}_1^+, \mathbf{v}_2^+, \ldots, \mathbf{v}_m^+\}$ and $\mathcal{S}^- = \{\mathbf{v}_1^-, \mathbf{v}_2^-, \ldots, \mathbf{v}_n^-\}$ be two sets of periodical vectors, the centroid of each set is computed as $\overline{\mathbf{v}}^+ = \frac{1}{m}\sum_1^m \mathbf{v}_i^+$ and $\overline{\mathbf{v}}^- = \frac{1}{n}\sum_1^n \mathbf{v}_j^-$. Then, the axis vector is defined as $\mathbf{v}_{\text{axis}} = \overline{\mathbf{v}}^+ - \overline{\mathbf{v}}^-$. We measure the projection of a periodical p to this axis using the cosine similarity between two vectors: $s(p, \mathbf{v}_{\text{axis}}) = \frac{\mathbf{v}(p) \cdot \mathbf{v}_{\text{axis}}}{|\mathbf{v}(p)| \cdot |\mathbf{v}_{\text{axis}}|}$. Here, we examine two spectra of scholarship: (i) "soft" to "hard" sciences (*35*–*37*) and (ii) social sciences to life sciences.

The first axis (dimension) captures the idea of the hierarchy of the sciences—an ordering of scientific disciplines by the complexity of the subject matter and the hypothesized order of development—which places natural sciences like Mathematics and Physics at the bottom and social sciences like Sociology at the top (*38*–*40*). Disciplines at the top of the hierarchy are argued to be soft—more complex, difficult to develop, and having less codified knowledge with more competing theories than disciplines at the bottom (*37*, *39*, *41*).

We operationalize the axis from soft to hard sciences using two sets of periodicals. The pole of the hard sciences is defined by the centroid of all journals in "Math & Physics" and the pole of soft sciences is defined by the centroid of all journals in "Social Sciences" and "Humanities" (table S1). We project each periodical p onto $\mathbf{v}_{\text{soft} \to \text{hard}}$ by calculating the cosine similarity $s(p, \mathbf{v}_{\text{soft} \to \text{hard}})$. The projection in Fig. 4A forms a continuous spectrum along this axis, documenting how academic journals are distributed along the given axis that runs from Social Sciences & Humanities to Mathematics & Physics.

Some exemplary hard journals include *Biophysical Journal*, *Journal of Theoretical Biology*, *Fractals*, *Physics Reports*, and *Physical Review E*. Some exemplary soft journals include *Applied Psychology*, *Anthropological Quarterly*, *Law & Society Review*, *Sociological Forum*, and *Politics & Society*. Several representative periodicals are annotated in the spectrum. We also rank 13 disciplines by the mean projection value of all journals in each category in Fig. 4A. The breakdown into each discipline provides richer insights into how major scientific branches are organized along this conceptual dimension (figs. S10 to S12). Overall, this spectrum shows that the "hardness" of academic disciplines increases in the order of Sociology, Psychology, Biology, Chemistry, Physics, and Mathematics, which concurs with the common conceptual ordering based on the hierarchy of the sciences (*38*, *39*, *42*).

The second dimension we examine is the one from social sciences to life sciences, another major branch of natural sciences. We place all "Social Sciences" and "Humanities" journals into the social sciences group, and all journals that are classified as "Biology," "Biotechnology," "Infectious Diseases," "Health Professionals," and "Medical Specialties" into the life sciences group. The spectrum of $\mathbf{v}_{\text{social} \to \text{life}}$ is shown in Fig. 4B. As expected, biomedical disciplines are located near the biological end of this spectrum. Most physical sciences, including "Chemistry," "Earth Sciences," and "Math & Physics", are distributed in the middle of this band. However, computer science, which was far from "Social Science" on the soft-hard sciences axis, is the closest to "Social Science" on this dimension. The same set of representative periodicals annotated in Fig. 4A is rearranged on the axis between social sciences and life sciences (Fig. 4B), highlighting the multifaceted nature of the disciplinary organization of periodicals and the embeddings' ability to tease out semantic dimensions.

The axis built by connecting the two centroids of two broad disciplines is robust to the selection of journals—a random sample of less than 1% journals in each pole discipline can well reproduce the ordering shown in Fig. 4. Furthermore, the overall axis created with this approach also correlates with that within the pole discipline built by using journals in their subdisciplines (see Materials and Methods).

DISCUSSION

Here, we present a continuous embedding framework for scholarly periodicals to systematically investigate the structure of periodicals and disciplines.

By applying our method to a large bibliographic dataset, we obtain continuous and dense vector representations of scientific periodicals that can better encode the relationships between periodicals than two citation-based sparse vector-space models. The periodical embeddings can also offer new measurements that overcome

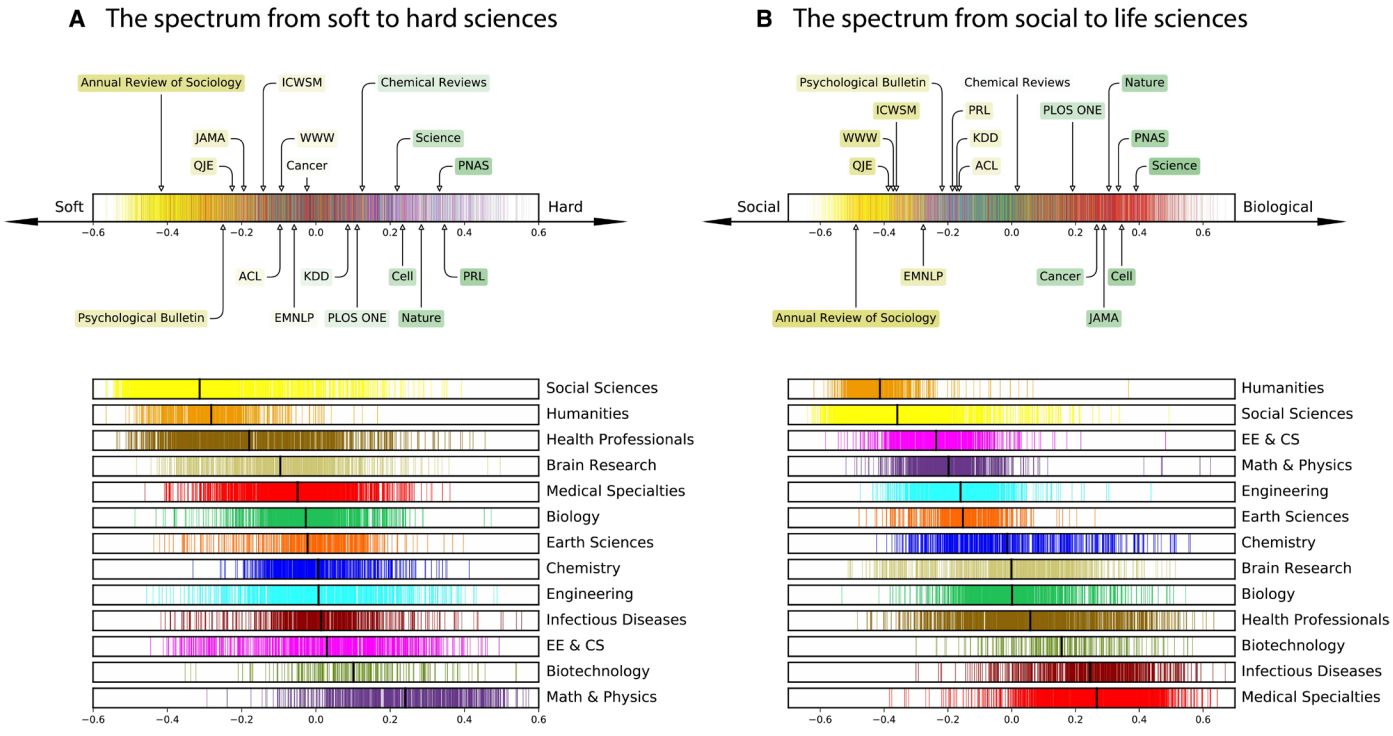

Fig. 4. Two spectra of scholarship. (**A**) Spectrum of soft and hard sciences, operationalized by defining $\mathcal{S}^{+} = \{\mathbf{v}(p) \mid p \in$ "Math & Physics"$\}$ and $\mathcal{S}^{-} = \{\mathbf{v}(p) \mid p \in$ "Social Sciences"$\vee\, p \in$ "Humanities"$\}$. Each disciplinary journal is represented by a vertical line inside the box (12,751 in total). The color represents the discipline category and the position reflects the cosine similarity between the periodical vector and the axis $\mathbf{v}_{\text{soft} \rightarrow \text{hard}}$. We also annotate several journals and proceedings, whose background colors are proportional to their projection values. We then show journals in each disciplinary category separately at the bottom. The black vertical line in each discipline represents the mean projection value of its journals. (**B**) The spectrum along the axis between social sciences and life sciences (biological), operationalized by defining $\mathcal{S}^{+} = \{\mathbf{v}(p) \mid p \in$ "Biology"$\vee p \in$ "Biotechnology"$\vee p \in$ "Infectious Diseases"$\vee p \in$ "Health Professionals"$\vee p \in$ "Medical Specialties"$\}$ and $\mathcal{S}^{-} = \{\mathbf{v}(p) \mid p \in$ "Social Sciences"$\vee p \in$ "Humanities"$\}$. Note that the ordering of 13 disciplines is dramatically changed from (A), reflecting the complex organization of scholarly periodicals in the embedding space along scientific axes.

conceptual and computational barriers. For instance, the framework allows us to make cross-disciplinary navigation using vector analogies and to organize periodicals and disciplines along conceptual scientific dimensions. More generally, the capacity to quantitatively operationalize relevant disciplinary dimensions will be useful for future studies that delve deeper into the complex disciplinary organization.

We acknowledge that there might exist comparable or better embedding methods given the rapid development in the field of machine learning (*43*). We view our primary contribution as one of the earliest attempts to apply the embedding approach to the science of science by (i) developing a method that is motivated by the citation flow between scientific periodicals and papers and (ii) extensively testing the usefulness of the embeddings with multiple evaluation tasks.

We also would like to point out limitations of our study. First, the quality of embeddings depends on the quality of the dataset; thus, our embeddings may reflect biases and errors in the data. For instance, it may be less useful for fields that are not covered by the source bibliometric dataset well. Second, the embedding approach suffers from sparse data; periodicals with fewer papers and citations will have embeddings that are less stable and less accurate, although it would still be better than only using explicit, direct links.

Third, as our method filters periodicals based on frequency (see Materials and Methods), it is possible that certain fields have fewer periodicals included in the embedding space. However, our matched 12,780 journals between MAG and UCSD catalog (table S1) indicate that every discipline still has at least several hundreds journals for the analyses, and the differences in coverage mainly result from their sizes according to the catalog (e.g., "Biotechnology" has only 11 subfields, but "Social Science" has 69 subfields in the UCSD catalog). Fourth, the embedding approach's assumption that vector representations are sufficient to capture explicit and implicit relationships between entities may not be valid. It has been argued that these embeddings may be impossible to obtain for networks with high clustering (*44*), although this may happen only in some embedding methods (*45*). Similarly, when there exist multiple contexts for each entity (*46*, *47*), a single vector may not be able to fully capture them. In our case, the embeddings of multidisciplinary journals may be skewed toward the primary disciplines that they publish and fail to capture explicit relationship to marginally published fields (fig. S3A). Thus, when it is critical to consider explicit connections, the embedding approach may be inappropriate. Furthermore, depending on the task, much simpler methods may well outperform embedding-based methods. For instance, a simple majority-voting approach outperforms our embedding in predicting the publication venue of a paper given its references (fig. S13).

Fifth, the present study does not take into account the evolution of periodicals and disciplines, falling short in providing a dynamic picture of the disciplinary patterns formed during different time periods. Sixth, although our method is more space and time efficient

than the sparse vector models in downstream analyses, it does require nonnegligible time and memory to train (e.g., a few hours to train the embeddings with 100 million citation trails), and thus, applying the method to a larger dataset such as the universe of scientific papers can be challenging. Last, because the exact mechanisms and properties of neural embedding methods have not been fully understood, there may be unknown biases in the periodical embeddings.

Despite these limitations, by demonstrating its validity and performance, we show that the embedding approach offers a promising avenue for science of science research.

Future work may extend our framework to develop better embedding methods, investigate fundamental scientific questions with new capacities offered by the embeddings, or model the evolution of scientific periodicals and disciplines by incorporating temporal information in citations.

MATERIALS AND METHODS

Dataset

We used the MAG data, which are the largest open access bibliometric dataset (*48*, *49*). The snapshot we used contains 126,909,021 papers published in 23,404 journals and 1283 conference proceedings between 1800 and 2016 (accessed on 5 February 2016). There are 528,245,433 citations between these papers.

We focused on all papers that were published in either journals or conference proceedings, as the periodical information is needed to train the embedding model. Thus, our study is based on a total number of 53,410,055 papers and 402,395,790 citations.

They were published between 1800 and 2016 in 24,020 scholarly periodicals. Figure S1 shows the number of papers over time.

Using our method, we obtained embeddings for 20,835 periodicals (3185 periodicals were dropped because of data filtering; see the "Hyperparameter tuning" section). However, MAG does not have the discipline information for these periodicals. We thus used the UCSD map of science catalog data (*9*), which contain discipline information for about 25,000 journals (classified into 13 academic disciplines). We matched 14,113 journals between MAG and the UCSD map on the basis of journal names, among which 12,780 journals are covered in our embeddings (table S1).

Model

We consider the citation network between papers, where each node is a paper and a directed edge from A to B is formed if paper A cites paper B. We generate many citation trails $\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_N\}$ from the citation graph using random walks, where we first randomly choose a starting point (a paper) and randomly follow citations until we arrive at a dead end (a paper without outgoing edges). Each trail $\mathcal{T}$ is a sequence of papers $(P_1^{\mathcal{T}}, P_2^{\mathcal{T}}, \ldots, P_{|\mathcal{T}|}^{\mathcal{T}})$. We discard trails that are immediately terminated ($|\mathcal{T}| = 1$). We then create a corresponding periodical trail $\mathcal{V}_{\mathcal{T}} = (V_1^{\mathcal{T}}, V_2^{\mathcal{T}}, \ldots, V_{|\mathcal{T}|}^{\mathcal{T}})$ for each paper citation trail, where the ith element $V_i^{\mathcal{T}}$ is the publication venue (periodical) of the ith paper $P_i^{\mathcal{T}}$ in the paper citation trail. Using the periodical trails, we learn two vector representations of each periodical $\mathbf{v}(V)$ ("input") and $\mathbf{v}'(V)$ ("output") by using the skip-gram with negative sampling (SGNS) method (*17*). For a given periodical citation trail $\mathcal{V}_{\mathcal{T}}$, the objective is to maximize the log probability

$$O = \frac{1}{|\mathcal{V}_{\mathcal{T}}|} \sum_{t=1}^{|\mathcal{V}_{\mathcal{T}}|} \sum_{-w \le j \le w, j \ne 0} \log p(V_{t+j}^{\mathcal{T}} \mid V_t^{\mathcal{T}}) \quad (1)$$

where w is the context window size. This training objective can be efficiently approximated as

$$E = \log \sigma(\mathbf{v}'(V_{\mathrm{O}})^{\top} \mathbf{v}(V_{\mathrm{I}})) + \sum_{i=1}^{k} \mathbb{E}_{V_i \sim \mathcal{U}(V)} [\log \sigma(-\mathbf{v}'(V_{\mathrm{i}})^{\top} \mathbf{v}(V_I))] \quad (2)$$

where V_{I} is the input periodical and V_{O} is the output (context) periodical in Eq. 1, and $\sigma(x) = 1/[1 + \exp(-x)]$. For each periodical pair $(V_{\mathrm{I}}, V_{\mathrm{O}})$, SGNS samples k negative pairs $(V_{\mathrm{I}}, V_{\mathrm{i}})$ from the empirical distribution $\mathcal{U}(V)$. Here, we let $k = 5$ and $\mathcal{U}(V)$ be the smoothed unigram distribution (*17*). After training, the input vectors are used as the periodical embeddings (*16*). All models are trained with $N = 100{,}000{,}000$. See the next section for details. SGNS method is efficient and scalable. It takes about 3 hours to train the embeddings with 100 million citation trails on a reasonably powerful computing server. The algorithm for training embeddings is implemented in the Gensim package (*50*).

Hyperparameter tuning

We tuned two hyperparameters of the SGNS model: the context window size (W) and the number of dimensions (D). For each combination of W (2, 5, 10, and) and D (50, 100, 200, and 300), we trained a model using the same 100 million periodical trails (fig. S2). We set the minimum periodical frequency to 50, which means that the embedding model will exclude periodicals with less than 50 occurrences because of data sparsity. A good model would output similar embedding vectors for periodicals that are similar in terms of research topics. We thus compared the quality of different embeddings on the basis of the cosine similarities between periodicals.

Specifically, we randomly sampled 100,000 journal pairs for each of the three groups: (i) in the same discipline, (ii) in the same subdiscipline, and (iii) random pairs. Note that we focused on 12,780 journals for which we have discipline categories and are covered in our embedding model (table S1). Table S2 indicates that the model trained with $W = 10$ and $D = 100$, which covers 20,835 periodicals, gives the best result. Figure 1C and fig. S3 (C and D) show that, based on the best model, journal pairs in the same discipline (and subdiscipline) are much more similar in the embedding space than those selected randomly from any discipline.

Journal recommendation survey

As an external evaluation, we use a survey to evaluate how well our embedding captures the similarity between periodicals. We focus on the 12,780 journals that have discipline categories (table S1) because some baselines rely on disciplinary classification. Each algorithm can rank, for a given target journal, the remaining 12,779 candidates. Note that the first baseline (disc.) gives an arbitrary rank for journals whose disciplines are different from that of the target. We designed a survey to evaluate the three algorithms and recruited faculty members, researchers, and doctoral students from University of Michigan and Indiana University (The University of Michigan institutional review board guidelines were followed with human subjects). To make the task feasible, we selected top 20 journals in each discipline based on their PageRank scores. Journals belonging to the "Interdiscipline" category were excluded in the survey. For each of the 260 target journals, we constructed a set of candidate journals and asked participants to rank them based on their topical similarities to the target. The candidate set is the union of the top four similar journals given by each algorithm. Because of

the overlap between the three top lists, the size of the candidate set varies between 4 and 12.

Participants first selected a discipline as their fields to begin the survey (fig. S4A). They were then asked about their familiarity with the 20 target journals in the selected discipline. Participants were allowed to continue the task only if they were familiar with at least three target journals (fig. S4B). Participants who selected less than three targets were immediately directed to the end of the survey. After the screening phase, the participants were asked to rank, for each selected target, the set of candidate journals on the basis of their topical similarities to the target. Participants can place unfamiliar candidates in the "Unfamiliar Journals" group (fig. S5).

Among 247 participants (of 367) who finished the survey, 119 were qualified to complete the ranking task, and each of them was rewarded a $10 Amazon gift card. Table S3 shows the statistics of qualified responses across different disciplines.

Experts could give quite different ranking of the same target journal. We used Kendall's rank correlation coefficient τ to measure the level of agreement between two ranked lists of a target, based on the intersection of two ranked lists. We focused on target journals $\mathcal{J}$ whose average pairwise expert agreement $\hat{\tau} \geq 0.2$. Note that this threshold is for determining which target journals should be used to evaluate the three algorithms. In the evaluation step, to leverage more expert information, we appended to each ranked list the unfamiliar journals in a random order (the results are qualitatively the same without including unfamiliar journals). Then, each ranked list for a target in $\mathcal{J}$ was used as the reference to evaluate three algorithms. Specifically, for a ranked list l_e^j of target journal j from an expert e, we retrieved, from the full ranked list of an algorithm a, the order l_a^j of journals in l_e^j, and we calculated $\tau_{(l_e^j, l_a^j)}$.

The average correlation between each algorithm and domain experts (Fig. 1D and fig. S6) indicates that, across three disciplines, the three vector-space models are better than the first baseline and are comparable to each other. The correlations between algorithms and experts are slightly higher with a higher threshold $\hat{\tau}$, but the error bars are also larger such that there is no clear winner between the three vector-space models (e.g., there are only two target journals with a total of four ranked lists for the evaluation with $\hat{\tau} = 0.8$).

Evaluation with the UCSD categorization

We test our hypothesis on the relationship between disagreement (with the UCSD catalog) and interdisciplinarity through a manual evaluation. First, we randomly selected 20 journals from the top 100 and the bottom 100 journals (10 from each) for three disciplines (EE & CS, Engineering, and Social Sciences) based on the agreement score obtained by comparing the UCSD catalog with a clustering produced by our embeddings using the element-centric similarity measure (*33*). We presented them to three of the authors and asked them to evaluate whether each journal belongs to the discipline defined in the UCSD categorization. Each person was given the following instruction: "Go to the journal's description page (via Google search or Wikipedia). Assign [yes] if only the target discipline is mentioned; Assign [no] if the target discipline is not mentioned; Assign [interdiscipline] if the target and other disciplines are mentioned; Assign [unsure] if no relevant information is found for this journal."

In the pretest, the average pairwise agreement for 60 journals in three disciplines was 0.59 (Cohen's kappa), which is moderately high (*51*). In the posttest, we again asked the three authors to evaluate another 20 journals for each of the three disciplines (each author evaluated a different set of journals). We combined the pretest responses (used the majority voting; 10 in the top and 10 in the bottom) and the posttest responses (30 in the top and 30 in the bottom) for each discipline. Journals with an [unsure] response were excluded in the analysis. We considered [yes] and [interdiscipline] as a journal being consistent with the UCSD catalog.

Validating the two spectra of science

We validate the spectrum of science in two ways. First, to test the robustness of the two dimensions, we rebuild the axis vector by connecting the centroids of a subset of randomly selected journals in the two pole disciplines and reorder all periodicals on the new axis. We then correlate this new ordering with their original arrangement (Fig. 4). We repeat this process 100 times. Figure S9 shows the average Spearman's rank correlation as a function of the number of journals used in the subset. The correlation is above 0.9 even when the new axis is built with less than 1% of all journals in the field for both the "soft-hard" axis and the "social-bio" axis. This high correlation suggests that the two axes are stable and robust.

Second, in our spectrum analysis, we used an axis anchored between the two centroids of two broad fields to score the fields themselves. We note that this could be problematic if the conceptual axis within the anchor field is not necessarily aligned with the overall axis, especially when the field does exhibit such an axis internally, such as Social Sciences (figs. S10 and S11).

To address this concern, we test whether the spectrum calculated at the level of the whole space is consistent with the spectrum calculated within the pole (anchor) discipline. Because it is unclear which subdisciplines of "Math & Physics" are "softer" or "harder" and the same issue seems to be true for Life Sciences with respect to the social-bio axis (figs. S10 and S11), we focus our efforts on social sciences.

To validate the soft-hard axis in ["Social Sciences" and "Humanities"], we first use "Sociology" as the soft subfield and "Finance"—which has close connections to mathematics and physics—as the hard subfield. We rebuild the axis vector by connecting the two centroids and reordered all 20,835 periodicals on this new axis. Although we use only two subdisciplines within a single discipline to obtain scores for all periodicals across all disciplines, the Spearman rank correlation between the new ordering and the original one (Fig. 4) is 0.73. This result is also robust. We construct nine soft-hard subfield pairs between three soft subdisciplines ("Law", "Social Psychology", and "Leadership & Organizational Behavior") and three hard subdisciplines ("Finance," "Statistics," and "Operations Research"). We then calculate the rank correlation between the ordering based on each pair and the overall ranking. The average correlation is 0.73 (95% confidence interval: [0.69, 0.77]).

We repeat this robustness test for the social-bio axis in ["Social Sciences" and "Humanities"], and we use "Sociology" as the "social" subfield and "BioStatistics" as the "bio" subfield. The Spearman rank correlations between the new ordering and the original one is 0.82. Similarly, the result is robust with other choices of subfields. We construct nine social-bio subfield pairs between three social subdisciplines ("Law," "Sociology," and "Economics") and three biological subdisciplines ("Psychiatric & Behavioral Genetics," "Psychosomatic Medicine," and "BioStatistics"). We then calculate the rank correlation between the ordering based on each pair and the overall ranking. The average correlation is 0.71 (95% confidence interval: [0.63, 0.78]).

Together, these results demonstrate that the conceptual axis (either soft-hard or social-bio) within the anchor fields is well aligned with the spectrum calculated with two broad disciplines, providing evidence that the two spectra are robust and meaningful.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/7/17/eabb9004/DC1

REFERENCES AND NOTES

1. E. Garfield, The history and meaning of the journal impact factor. *JAMA* **2950**, 90–93 (2006).
2. A. Fersht, The most influential journals: Impact factor and eigenfactor. *Proc. Natl. Acad. Sci. U.S.A.* **1060**, 6883–6884 (2009).
3. M. Baldwin, *Making "Nature": The History of a Scientific Journal* (University of Chicago Press, 2015).
4. A. Csiszar, *The Scientific Journal: Authorship and the Politics of Knowledge in the Nineteenth Century* (University of Chicago Press, 2018).
5. R. K. Merton, *The Sociology of Science: Theoretical and Empirical Investigations* (University of Chicago Press, 1973).
6. K. W. Boyack, R. Klavans, K. Börner, Mapping the backbone of science. *Scientometrics* **64**, 351–374 (2005).
7. M. Rosvall, C. T. Bergstrom, Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1118–1123 (2008).
8. J. Bollen, H. Van de Sompel, A. Hagberg, L. Bettencourt, R. Chute, M. A. Rodriguez, L. Balakireva, Clickstream data yields high-resolution maps of science. *PLOS ONE* **4**, e4803 (2009).
9. K. Börner, R. Klavans, M. Patek, A. M. Zoss, J. R. Biberstine, R. P. Light, V. Larivière, K. W. Boyack, Design and update of a classification system: The UCSD map of science. *PLOS ONE* **7**, e39464 (2012).
10. B. Uzzi, S. Mukherjee, M. Stringer, B. Jones, Atypical combinations and scientific impact. *Science* **342**, 468–472 (2013).
11. H. Small, Visualizing science by citation mapping. *J. Am. Soc. Inf. Sci.* **50**, 799–813 (1999).
12. K. Börner, *Atlas of Science: Visualizing What We Know* (The MIT Press, 2010).
13. G. Salton, A. Wong, C.-S. Yang, A vector space model for automatic indexing. *Commun. ACM* **18**, 613–620 (1975).
14. Y. L. Cun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
15. O. Levy, Y. Goldberg, I. Dagan, Improving distributional similarity with lessons learned from word embeddings. *Trans. Assoc. Comput.* **3**, 211–225 (2015).
16. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs.CL] (16 January 2013).
17. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in Proceedings of the 26th International Conference on Neural Information Processing Systems (2013), pp. 3111–3119.
18. A. Mnih, K. Kavukcuoglu, Learning word embeddings efficiently with noise-contrastive estimation, in Proceedings of the 26th International Conference on Neural Information Processing Systems (2013), pp. 2265–2273.
19. Y. Dong, N. V. Chawla, A. Swami, metapath2vec: Scalable representation learning for heterogeneous networks, in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2017), pp. 135–144.
20. J. An, H. Kwak, Y.-Y. Ahn, Semaxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (2018), pp. 2450–2461.
21. W. L. Hamilton, J. Leskovec, D. Jurafsky, Diachronic word embeddings reveal statistical laws of semantic change, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (2016), pp. 1489–1501.
22. M. Rudolph, D. Blei, Dynamic embeddings for language evolution, in Proceedings of the 2018 World Wide Web Conference (2018), pp. 1003–1011.
23. T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, in Proceedings of the 30th International Conference on Neural Information Processing SystemsDecember (2016), pp. 4356–4364.
24. N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E3635–E3644 (2018).
25. A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
26. A. C. Kozlowski, M. Taddy, J. A. Evans, The geometry of culture: Analyzing meaning through word embeddings. *Am. Sociol. Rev.* **84**, 905–949 (2018).
27. Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
28. Q. Le, T. Mikolov, Distributed representations of sentences and documents, in Proceedings of the 31st International Conference on Machine Learning (2014), pp. 1188–1196.
29. T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 4401–4410.
30. B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: Online learning of social representations, in Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2014), pp. 701–710.
31. A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016), pp. 855–864.
32. M. Rosvall, A. V. Esquivel, A. Lancichinetti, J. D. West, R. Lambiotte, Memory in network flows and its effects on spreading dynamics and community detection. *Nat. Commun.* **5**, 4630 (2014).
33. A. J. Gates, I. B. Wood, W. P. Hetrick, Y.-Y. Ahn, Element-centric clustering comparison unifies overlaps and hierarchy. *Sci. Rep.* **9**, 8574 (2019).
34. M. Nissim, R. van Noord, R. van der Goot, Fair is better than sensational: Man is to doctor as woman is to doctor. *Comput. Linguist.* **46**, 487–497 (2020).
35. In praise of soft science. *Nature* **435**, 1003 (2005).
36. S. Cole, Why sociology doesn't make progress like the natural sciences. *Sociol. Forum* **9**, 133–154 (1994).
37. L. V. Hedges, How hard is hard science, how soft is soft science? the empirical cumulativeness of research. *Am. Psychol.* **42**, 443–455 (1987).
38. A. Comte, *The Positive Philosophy of Auguste Comte* (Calvin Blanchard, 1855).
39. S. Cole, The hierarchy of the sciences? *Am. J. Sociol.* **89**, 111–139 (1983).
40. D. Fanelli, W. Glänzel, Bibliometric evidence for a hierarchy of the sciences. *PLOS ONE* **8**, e66938 (2013).
41. J. B. Lodahl, G. Gordon, The structure of scientific fields and the functioning of university graduate departments. *Am. Sociol. Rev.* **37**, 57–72 (1972).
42. R. Munroe, Purity. *xkcd.com* (2019); https://xkcd.com/435/.
43. P. Goyal, E. Ferrara, Graph embedding techniques, applications, and performance: A survey. *Knowl.-Based Syst.* **151**, 78–94 (2018).
44. C. Seshadhri, A. Sharma, A. Stolman, A. Goel, The impossibility of low-rank representations for triangle-rich complex networks. *Proc. Natl. Acad. Sci. U.S.A.* **1170**, 5631–5637 (2020).
45. S. Chanpuriya, C. Musco, K. Sotiropoulos, C. E. Tsourakakis, Node embeddings and exact low-rank representations of complex networks. arXiv:2006.05592 [cs.LG] (10 June 2020).
46. G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).
47. Y.-Y. Ahn, J. P. Bagrow, S. Lehmann, Link communities reveal multiscale complexity in networks. *Nature* **466**, 761–764 (2010).
48. A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, K. Wang, An overview of microsoft academic service (mas) and applications, in Proceedings of the 24th International Conference on World Wide Web (2015), pp. 243–246.
49. B. K. AlShebli, T. Rahwan, W. L. Woon, The preeminence of ethnic diversity in scientific collaboration. *Nat. Commun.* **9**, 5163 (2018).
50. R. Řehůřek, P. Sojka, Software framework for topic modelling with large corpora, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (2010), pp. 45–50.
51. M. L. McHugh, Interrater reliability: The kappa statistic. *Biochem. Med.* **22**, 276–282 (2012).
52. L. van der Maaten, G. Hinton, Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

Acknowledgments: We thank C. Quarles, X. Yan, C. R. Sugimoto, and V. Larivière for helpful discussion. **Funding:** This work is supported, in part, by the Air Force Office of Scientific Research under award numbers FA9550-19-1-0391 and FA9550-19-1-0029. **Author contributions:** All authors designed the study. H.P. performed the analyses. H.P. and Y.-Y.A. produced the figures. H.P. and Y.-Y.A. led the writing of the manuscript. All authors contributed to the writing and approved the final manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. All code used in this study is available at: https://github.com/haoopeng/periodicals. The MAG data can be accessed at: https://microsoft.com/en-us/research/project/microsoft-academic-graph/. A public repository of our data is available at: https://doi.org/10.6084/m9.figshare.13007650. Additional data related to this paper may be requested from the authors.

Submitted 25 March 2020
Accepted 26 February 2021
Published 23 April 2021
10.1126/sciadv.abb9004

Citation: H. Peng, Q. Ke, C. Budak, D. M. Romero, Y.-Y. Ahn, Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *Sci. Adv.* **7**, eabb9004 (2021).

Neural embeddings of scholarly periodicals reveal complex disciplinary organizations

Hao Peng, Qing Ke, Ceren Budak, Daniel M. Romero and Yong-Yeol Ahn

Sci Adv **7** (17), eabb9004.
DOI: 10.1126/sciadv.abb9004