[9]7/10 2019-20pages- 2021.11.01-11.03

PROCEEDINGS A

[TO BE COMPLETED]

royalsocietypublishing.org/journal/rspa

→ interesting framework
network analysis methodology]
→ Need to look into more details
about adaptation

Research

Cite this article: Christianson NH, Sizemore Blevins A, Bassett DS. 2020 Architecture and evolution of semantic networks in mathematics texts. *Proc. R. Soc. A* **476**: 20190741.
http://dx.doi.org/10.1098/rspa.2019.0741

Received: 31 October 2019
Accepted: 5 June 2020

Subject Areas:
pattern recognition, complexity, applied mathematics

Keywords:
mathematics education, network science, persistent homology, semantic networks, learning

Author for correspondence:
Danielle S. Bassett
e-mail: dsb@seas.upenn.edu

One contribution to a special feature 'A generation of network science'.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5056734.

THE ROYAL SOCIETY
PUBLISHING

Architecture and evolution of semantic networks in mathematics texts

Nicolas H. Christianson[1,2], Ann Sizemore Blevins[2] and Danielle S. Bassett[2,3,4,5,6,7]

[1]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
[2]Department of Bioengineering, School of Engineering and Applied Science, [3]Department of Physics and Astronomy, College of Arts and Sciences, [4]Department of Electrical and Systems Engineering, School of Engineering and Applied Science, [5]Department of Neurology, Perelman School of Medicine, and [6]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[7]Santa Fe Institute, Santa Fe, NM 87501, USA

NHC, 0000-0001-8330-8964; DSB, 0000-0002-6183-4493

Knowledge is a network of interconnected concepts. Yet, precisely how the topological structure of knowledge constrains its acquisition remains unknown, hampering the development of learning enhancement strategies. Here, we study the topological structure of semantic networks reflecting mathematical concepts and their relations in college-level linear algebra texts. We hypothesize that these networks will exhibit structural order, reflecting the logical sequence of topics that ensures accessibility. We find that the networks exhibit strong core–periphery architecture, where a dense core of concepts presented early is complemented with a sparse periphery presented evenly throughout the exposition; the latter is composed of many small modules each reflecting more narrow domains. Using tools from applied topology, we find that the expositional evolution of the semantic networks produces and subsequently fills knowledge gaps, and that the density of these gaps tracks negatively with community ratings of each textbook.

Broadly, our study lays the groundwork for future efforts developing optimal design principles for textbook exposition and teaching in a classroom setting.

1. Introduction

Knowledge has been distilled into formal representations for millennia [1,2]. Such efforts have sought to explain human reasoning and support artificial intelligence [3–5]. Semantic networks organize information by detailing concepts and their relations as the nodes and edges of a graph [6]. In an educational context, concept maps reflect students' understanding of information in a similar manner, but may be used to evaluate comprehension [7–10] and identify topics that are most difficult to connect to other concepts [11]. With the capacity to construct semantic networks, concept maps and similar formal representations of knowledge comes the challenge of distilling mechanisms of knowledge acquisition.

Network science offers an appropriate conceptual language and useful mathematical toolset with which to meet this challenge [12]. In the parlance of network science, semantic networks of language tend to exhibit highly ordered architectures with strong local clustering, relatively short paths between any pair of nodes, and a few hubs, which are connected to an unexpectedly large number of other nodes [6,13]. Recent work using highly stylized laboratory experiments provides some preliminary evidence that network structure may play a role in how humans process information [14–16] and acquire knowledge [17–19]. Yet extending these findings to the real world has proven difficult, and it remains unknown precisely how the network structure of knowledge in the form of science textbooks [20], science and mathematics topics on Wikipedia [21], and even formal scientific papers [22,23] impacts the learnability of these content domains. Furthermore, as learning is a process, studies of semantic network architecture would benefit from evaluating a network's dynamic structure as it unfurls over the course of presentation, exposition or acquisition. The education literature establishes that the order in which topics are introduced can help or hinder learning at this level [24,25], but a rigorous understanding of order and dynamic structure in knowledge acquisition has not been formalized in ecologically valid experimental settings.

Here we seek to address these limitations by studying semantic networks of mathematical concepts in linear algebra textbooks [26,27]. A common college-level course, the subject is rigorous and logical, sequentially introducing concepts that naturally relate to, depend on and follow from other concepts. To begin, we seek to understand the structure of these inter-concept relations in textbooks, which present the knowledge in a thoughtfully ordered and comprehensive exposition. While each author may introduce and relate topics in a different order, we assume that each text serves to elucidate and approximate the latent structure of the domain of knowledge it conveys. Using techniques from network science [12], we test the hypothesis that these semantic networks exhibit structural order, indicating a logical sequence of topics that ensures accessibility. Motivated by a recent report that language acquisition proceeds through an ordered progression filling knowledge gaps [28], we use *persistent homology* [29–32] to track the growth and development of topological cavities in the semantic network. We predict that fewer knowledge gaps will exist in the texts than in null models of randomly growing semantic networks; withholding connections between topics that have already been taught is unlikely to effectively convey knowledge. Finally, we compare the growth of semantic networks elicited from multiple texts, in terms of their different expositional structures and topic orderings. We hypothesize that the degree to which knowledge gaps are created and persist within texts may be related to the complexity or difficulty of a text, and to the knowledge it conveys. Broadly, our quantitative evaluation of the differing structures and expositional layouts of distinct textbooks provides a foundation for future work examining the effects of topic ordering and network architecture on classroom learning.

2. Results

We constructed semantic networks and expositional growing networks from 10 linear algebra textbooks (see Methods). We first used a modified version of the RAKE algorithm [33] to identify significant phrases (figure 1, step 1), which we refer to collectively as the index list of concepts. We represent these concepts as nodes, and connect two nodes by an edge if their corresponding concepts co-occur within the same sentence (figure 1, step 2). To mimic the growth of a reader's knowledge network, we add nodes and edges as soon as they are mentioned in the book (figure 1, step 3). Across textbooks, node sets ranged in length from 146 to 453 (average 279.4) and edge densities ranged from 0.0748 to 0.204 (average 0.129). In what follows, we characterize the semantic network growth of all texts, and when useful we give examples from individual texts referred to by author last name.

(a) Meso-scale structure of semantic networks

Mathematics as a field and linear algebra as a subject contain many fundamental topics and conceptual connections between those topics. Practitioners and authors might contest which topics are fundamental, and which are more tangential, or less strongly linked to the rest. Within a network, this organizational scheme can manifest as core–periphery structure where fundamental concepts are densely connected to one another, while peripheral concepts connect to the core but not to one another (figure 2*a*). To assess this structure in a semantic network constructed from the whole text, we calculate the core–periphery statistic and compare the statistic values to those obtained from two null models (figure 2*c*): (i) a *random index null model*, in which the RAKE concept extraction methodology is bypassed, and words that are not common 'stop words' (see electronic supplementary material, Methods) are chosen at random within each text to serve as the index sets for generating expositional networks, and (ii) a *continuous configuration model*, in which the original network is rewired while maintaining node degree and strength. Generally, we observe that the empirical semantic networks show greater core–periphery organization than the continuous configuration model, suggesting the presence of a strongly connected core of topics along with a set of sparsely connected periphery topics given the degree and strength distributions. Interestingly, we also observe that the empirical networks show less core–periphery organization than the random index model, indicating that the networks of math terms are more homogeneous than a network of randomly chosen words.

We next investigate the internal structure of the core and periphery. For the core, we find that across texts many similar words participate, including 'determinant', 'vector space' and 'matrix' as expected (see electronic supplementary material, table S3). By contrast, we expect that the periphery contains terms more specific to a given book and its particular sub-topics. We therefore hypothesize that the periphery will display community structure (figure 2*d*). To test our hypothesis, we calculate the modularity of the periphery subnetwork, along with the relevant subnetworks of the random index and continuous configuration null models. We observe that the periphery of each semantic network generally exhibits a modular organization that is stronger than that of the continuous configuration model, but weaker than that of the random index model (figure 2*e*). Intuitively, while randomly chosen words may display strong modularity due to greater variation in semantic relationships and frequencies, mathematics phrases are used in a more modular fashion than expected from the rewired continuous configuration model, perhaps due to the nature of focusing on one general idea at a time in chapters and sections.

(b) Expositional development of the large-scale structure

How does the identified network structure develop along a text's exposition? We find that the expositional introduction of nodes in the final network's core precedes the introduction of periphery nodes throughout the exposition (figure 3*a*). We quantify this observation by calculating the area between the core and periphery node introduction curves; high values

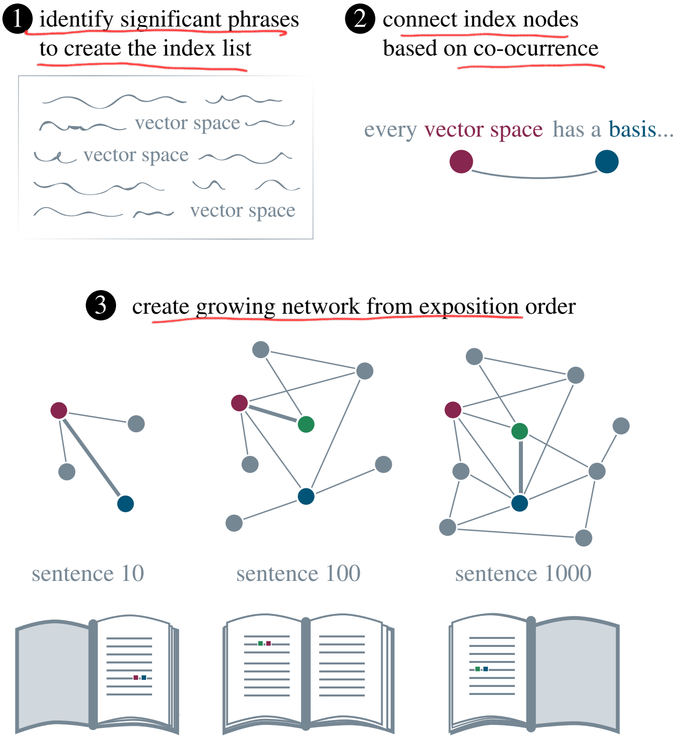
1 identify significant phrases to create the index list
vector space
vector space
vector space
2 connect index nodes based on co-ocurrence
every vector space has a basis...
3 create growing network from exposition order
sentence 10
sentence 100
sentence 1000

Figure 1. Extracting growing semantic networks from textbooks. (1) The index set is populated with phrases conveying significant mathematical concepts. (2) Any index nodes that co-occur within the same sentence are connected by an edge. (3) This procedure is applied to each sentence in the exposition, forming a view of the semantic network as it grows throughout the text. (Online version in colour.)

indicate that the core appears much earlier than the periphery, and low values indicate that the core and periphery appear at a more equal rate. The areas range from 0.064 and 0.20 across texts, and a one-sample t-test rejects the null hypothesis that these values are drawn from a distribution with mean 0 ($t = 8.65$, $p = 1.18 \times 10^{-5}$; calculated with the SciPy library, v. 1.1.0 [35]).

Next, we compare the areas obtained from the texts to those expected in statistical null models. Notably, we find that the empirical periphery is introduced earlier (relative to the core) than expected from the random index model, which has a more stark difference between core and periphery introduction (figure 3*b*). We observe no consistent trend across texts in comparison to a *random sentence model*, in which we use the original index list to build a growing graph from the texts after randomizing sentence order. While many texts show a marked discrepancy between the core and periphery development, others show a more even development. These differences across texts could reflect different expositional styles among different authors: some may choose to introduce core topics initially and save extra tangents for later, while others may involve discussions of peripheral topics throughout the text for motivation. Additionally, we take a similar approach in examining the relative rate of introduction of edges connecting different types of groups within the core and periphery, and find that of all edge types, those connecting concepts within a single periphery community are introduced the most sporadically, with some communities being fully introduced early on in the text, and some being introduced later (see electronic supplementary material, figure S2).

(c) Expositional development of knowledge gaps

In studying core and periphery formation, we focused on densely connected areas in the growing networks; now we turn to a study of sparsely connected areas. Specifically, we seek to understand

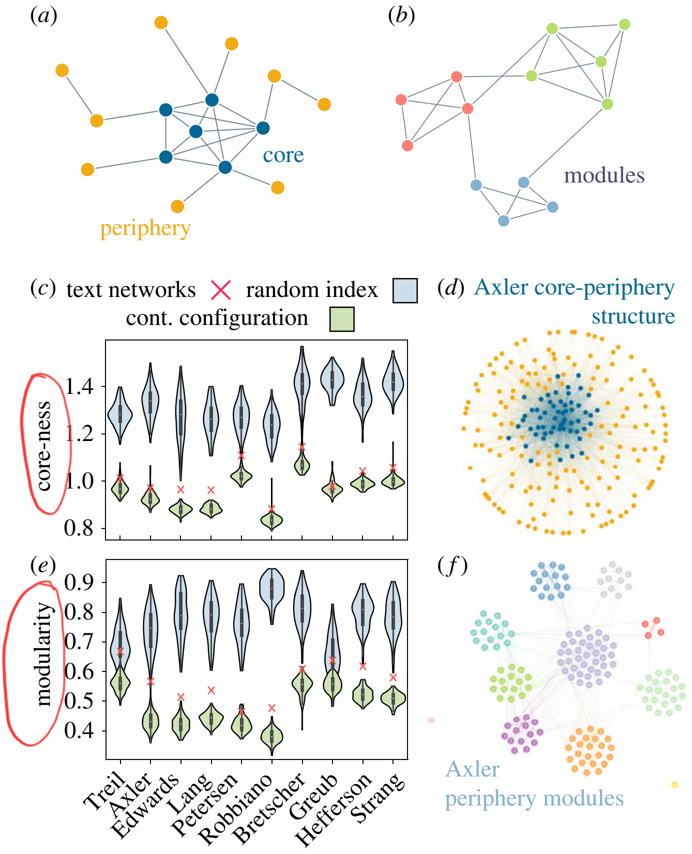

Figure 2. Meso-scale structure of semantic networks. (*a*) A schematic of core–periphery structure, with densely connected core nodes and a sparsely connected periphery. (*b*) A schematic of community structure, with densely connected communities which are themselves sparsely connected to each other. (*c*) The core-ness statistic of each network and of the corresponding random index and continuous configuration null ensembles. (*d*) Visualization of the Axler core–periphery structure. (*e*) Modularity statistics of the periphery of each network and of the corresponding random index and continuous configuration null ensembles. (*f*) Visualization of the Axler periphery community structure. Graph visualizations generated with graph-tool [34]. See electronic supplementary material, table S4, for example, nodes present in the Axler periphery communities. (Online version in colour.)

how voids or knowledge gaps might emerge and evolve throughout the exposition. Teaching strategies may intentionally leave open a connection or an area of the knowledge space in order to more intuitively reveal the connection later when a learner has more experience, or to provide the reader the opportunity to derive the connection on his/her/their own. A lack of connections between concepts can manifest as a topological gap in the network (figure 4*a*).

To detect gaps that form and evolve throughout the text, we compute the persistent homology [29–32] of the ordered set of networks composed of nodes and edges that exist at each point in the exposition; note, this ordered set of binary graphs is referred to as a filtration. We specifically detect gaps between connected components (dimension 0 homology > 1), cavities within rings of edges (dimension 1 homology > 0) or voids within polyhedra (dimension 2 homology > 0). We say that these so-called persistent cavities are *born* at the first instance of their appearance in the network, they *live* as long as the network grows and the topological void still persists, and they *die* when they are either connected to another previously disconnected component (in the case of dimension 0) or are tessellated by crossing edges (in the higher dimension cases). We invite the reader to refer to the Methods for a more rigorous description of persistent homology in this application.

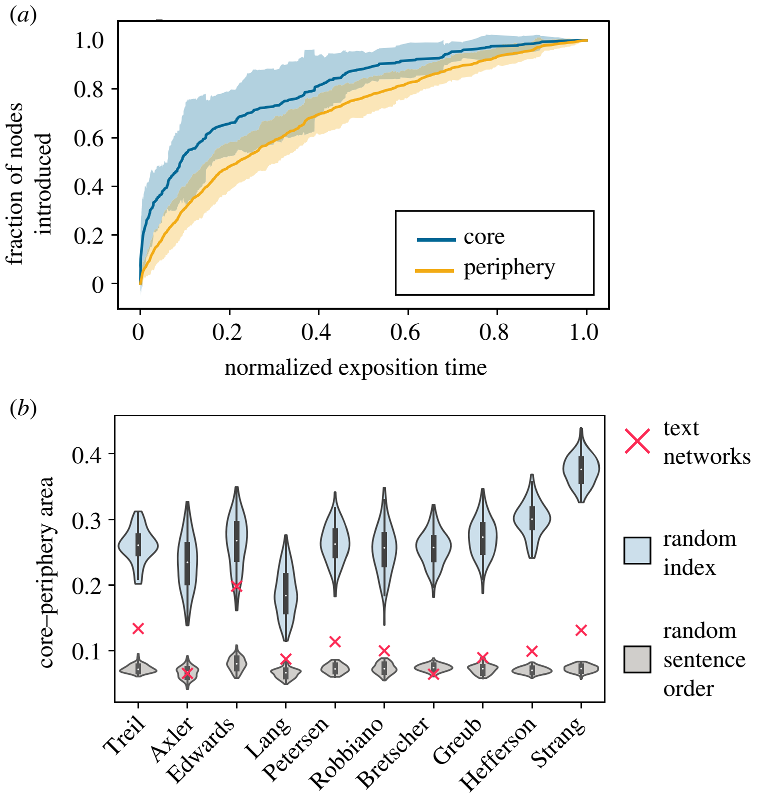

Figure 3. Development of core–periphery structure during exposition. (*a*) Core and periphery development curves, showing the fraction of nodes in each group introduced by a particular time in the exposition; mean ±2 standard deviations across all texts. (*b*) Difference in area between core and periphery development curves for all texts and corresponding random index and random sentence order null ensembles. (Online version in colour.)

In order to detect emerging and evolving gaps throughout exposition, we compute the persistent homology of each text. The number of gaps of dimension n that are alive at a given point in the filtration, called the Betti curve, is denoted β_n. We see that the texts tend to generate a large number of components, as manifest by the initial β_0 peak, followed by a rise in β_1, and finally a slow and steady increase in β_2 (figure 4*b*). For each text, we summarize the life and death of each persistent gap in a barcode (figure 4*c*). Each bar represents a single persistent cavity; the left endpoint of the bar indicates the birth time of the persistent cavity, while the right endpoint indicates the death time. Across all texts, we see that although many persistent cavities are killed soon after birth, a non-trivial number of gaps in each of the three dimensions persist throughout many sentences, suggesting that long-lived gaps are a consistent hallmark of the growing text structure.

To further evaluate the substantiveness of the gap architecture, we compared the persistent homology of the text networks to two filtration-based null models. In the first null model, we use the introduction of concepts to order the complete network for the filtration. More precisely, this node-ordered null model adds a node at the first mention of the concept, and also adds all of the connections that will ever exist from that node to previously acquired concepts. This model mimics the teaching strategy of introducing all connections of a new concept to anything previously taught. We find that the node-ordered model produces almost no persistent homology, in stark contrast to the original text (see electronic supplementary material, figures S5 and S6). This result suggests that the text expositions consistently leave connections between already-learned concepts for later discussion. We use a second null model to determine whether the totally

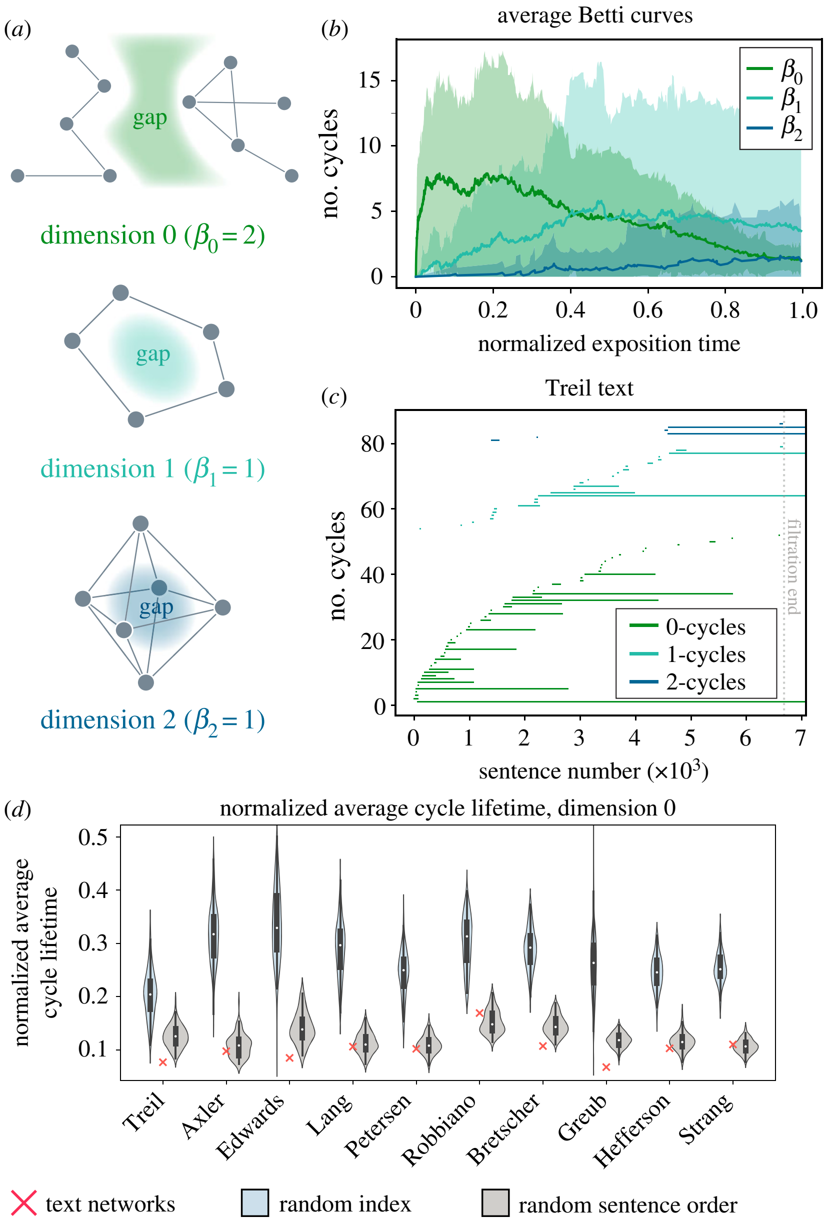

Figure 4. Development and persistence of knowledge gaps throughout exposition. (*a*) Examples of knowledge gaps in dimensions 0, 1 and 2. (*b*) Number of live knowledge gaps (β_n) in each dimension n throughout exposition; mean ± 2 standard deviations across all texts. (*c*) Barcode for the Treil text, showing introduction, persistence and death of cycles introduced throughout exposition. Barcodes for other texts are provided in electronic supplementary material, figures S3–S6. (*d*) The 0-dimensional normalized average cycle lifetime across all texts, as well as corresponding random index and random sentence order null models. (Online version in colour.)

random introduction of edges might produce similar progressions of persistent cavities. We find that this random edge order model produces an order of magnitude more persistent cavities of dimension 1 and 2 than the original text (see electronic supplementary material, figures S5 and S6). Broadly, the presence of a few long-lived cavities in the actual text are consistent with the notion that knowledge gaps exist but are introduced sparingly, and that introducing connections to all topics previously learned is not the strategy of these texts.

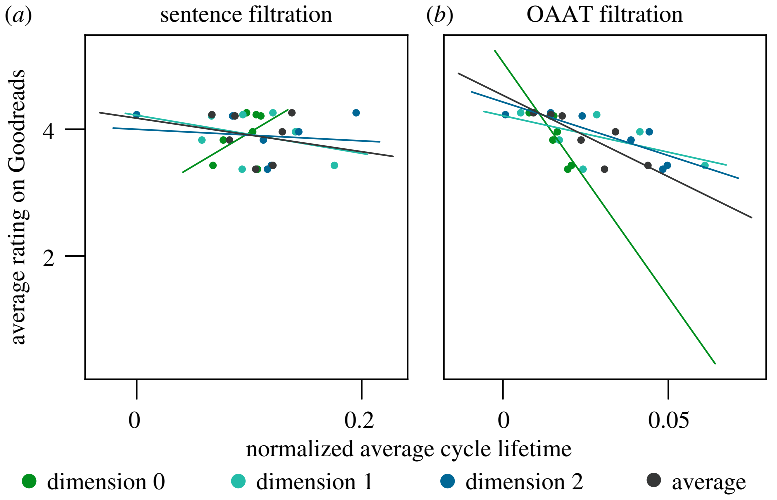

Figure 5. Relationship between textbook ratings and network topology. Scatterplots and best-fit lines for average Goodreads rating versus the normalized average cycle lifetime in dimensions 0 through 2 and the average over dimensions, across all texts for the (*a*) sentence filtrations and (*b*) OAAT filtrations. (Online version in colour.)

At this point, we know that throughout the text the introduction of terms and connections forms and fills gaps as a reader progresses. However, we do not yet know if the number and longevity of persistent cavities is different than we would expect from any growing semantic network in the text or from a reordered text. In order to answer this question, we define the normalized average cycle lifetime in dimension n as the sum of all persistent cavity lifetimes normalized by the number of cavities and filtration length (similar to metrics defined in [36], see Methods for details). Then intuitively a large value of normalized average cycle lifetime suggests that multiple long-lived persistent gaps exist, while a small value suggests that any gaps that form will die shortly after birth. We show the distributions of normalized average cycle lifetime values in dimension 0 in figure 4*d* for the random index and random sentence models, and the corresponding distributions for dimensions 1 and 2 in electronic supplementary material, figures S7 and S8. For completeness, we also include the barcodes and Betti curves for each model in the electronic supplementary material, figures S3–S6. Strikingly, the original text expositions generally fall below both null models' expected normalized average cycle lifetimes in dimension 0. This observation suggests that the exposition proceeds in a manner that may intentionally avoid developing disconnected topics, or possibly connects new topics to others very quickly. In dimensions 1 and 2, texts' normalized average cycle lifetimes vary more in relation to their null models, with only a handful of texts showing lower values than the null models.

(d) Evolving structure and text properties

After characterizing the structural features of the growing text networks, we next ask if these features might relate to text rating. Perhaps some readers particularly enjoy a book that leaves open many gaps motivating future study, while others enjoy a book with a stronger core offering conceptual closure. To determine whether readers' preferences relate to network structure, we used average text rating across all editions from Goodreads (goodreads.com). We kept any text which had at least five ratings, which was the case for seven of the 10 texts. We observe no significant correlation between average text rating and normalized average cycle lifetime across texts' sentence-based filtrations (figure 5*a*; see electronic supplementary material, table S5 for all Spearman's correlation coefficients and p-values). We also consider a *one-at-a-time* (OAAT) filtration (see electronic supplementary material, Methods), which in addition to allowing for comparability in persistent homology across texts and null models, provides additional information not just about a text's knowledge gaps on the sentence scale, but furthermore its sub-sentence topological structure. Remarkably, we observe significant negative correlations between average rating and OAAT normalized average cycle lifetime in dimensions 0 (Spearman's correlation coefficient $\rho = -0.857$, $p = 0.0137$) and 2 ($\rho = -0.893$, $p = 0.00681$), as well as the mean cycle lifetime averaged over dimensions 0, 1, and 2 ($\rho = -0.821$, $p = 0.0234$) (figure 5*b*, see electronic supplementary material, table S5 for all statistics). Intuitively, these results provide preliminary support for the notion that the extent of knowledge gaps in exposition influences the quality of a text as a learning tool. However, these results only account for seven of the 10 books, due to lack of availability of ratings for the others; as such, further work will be necessary to confirm the reliability of these findings in larger samples. For a description of additional relationships between the texts' structural features and broader text characteristics, we refer the reader to the electronic supplementary material.

3. Discussion

Here we examined the structure and topological development of semantic networks of mathematical knowledge as extracted from linear algebra texts. Meso-scale structural analysis indicates that the semantic networks exhibit strong core–periphery structure, where a tightly knit group of concepts form a core, surrounded by sparsely connected periphery concepts that are grouped into communities. Furthermore, these features appear to relate to the growth of the networks over the course of exposition; the cores of networks are built more quickly than the peripheries, and edges within each particular periphery community are introduced at varied times over the course of exposition. Using persistent homology, we extracted the knowledge gaps inherent in the exposition and found that the number of distinct connected components tends to decrease throughout the text, while topological cavities tend to increase. Finally, we examined possible relationships between the extent and persistence of knowledge gaps and other features of a text and its associated semantic network, providing motivation for future work examining the role of knowledge gaps in learning.

(a) Structure and evolution of meso-scale features in semantic networks

The prevalence of core–periphery and community structures in the networks we examine is consistent with a hierarchical structuring of mathematical knowledge, in which there exist a set of foundational concepts (the core), which are necessary for the subsequent logical development of subsidiary (periphery) concepts, which themselves are hierarchically organized into related communities. The generic notion of hierarchical structure in mathematics has been discussed in the context of presenting a logical sequence of concepts in education [37]. Hierarchical structure has also been noted in Wikipedia topic networks, in which concepts tend to maintain several connections to the foundational concepts used in each article [21]. Furthermore, hierarchical structure is observed in rigorous formulations of mathematics, which serves as the latent structure of mathematical knowledge. Formal proof networks for a variety of theorems across fields such as geometry and algebra exhibit great modularity and a degree distribution characteristic of a hierarchical, assemble-and-tinker construction process [38]; additionally, a hierarchical model of random Boolean networks recovers a similar bias-complexity relationship to that observed in formulas in propositional logic [39]. A hierarchical structure of mathematics knowledge is intuitive, particularly within a delimited area such as linear algebra: a set of foundational concepts, such as matrix, vector and linearity, is used to motivate and develop the rest of the topics within the field, which, for the most part, all presuppose the concepts in the core. Naturally, this hierarchy will not be a simple dichotomy (core–periphery), but the subsidiary concepts should themselves fall hierarchically into different groups, which may differ across texts due to author interests and publisher goals.

The observed growth dynamics of core–periphery structure offers a coherent expositional model. Given that the set of core concepts are highly related, and thus plausibly represent the concepts providing the foundation for linear algebra, it seems reasonable to introduce these concepts early and to introduce periphery concepts, which presuppose the core, later. The importance of giving sufficient foundational context and prior knowledge in exposition is well appreciated [40]. Further, the edge dynamics we observe are consistent with an expositional model in which topics are procedurally related to each other; the core concepts are introduced first, and used to introduce, at each point in the exposition, the communities that are being focused on; furthermore, subsidiary concepts that have already been introduced may then be used to give context for and develop further, separate subsidiary communities. Such an expository approach, in which connections are consistently introduced between that which has been learned, and that which is to be learned, has been demonstrated as useful in teaching mathematics proofs [41]. An extension of this motivating expository style is one that incorporates historical context [42]. Much prior work has sought to map the historical structure of research in scientific fields [43]. Recent work in particular has sought to track the temporal evolution of fields and their topics, for example, studying the exchange of ideas between topics in information retrieval [44], the community structure and prevalence over time of topics within neuroimaging [45] and the temporal dynamics of the interdisciplinarity of work across disparate fields of science [46]. In the future, it could prove fruitful to compare the expository structure of mathematics and other scientific texts with the historical development of the results included in those texts.

It is worth noting that, while recent efforts address the dynamics of core–periphery [47,48] and community [49,50] structures in networks, comparatively little work has addressed the growth and emergence of these structures over time. The perspective of our work is therefore important; we consider there to be some *a priori* structure of mathematical knowledge, unbeknownst to the reader, which each author seeks to convey. Thus, rather than examining the dynamics of meso-scale features in the networks, we instead focus on how the eventual features are created throughout exposition. Importantly, these eventual features may not exactly resemble those present in the latent structure of the mathematics itself, or in the author's understanding of the mathematics. In particular, our approach has been to examine the exposition of mathematics, whereas the rigorous, proof-based formulation of that mathematics may differ in form from its exposition. Nonetheless, we anticipate that insofar as mathematical exposition is meant to convey an understanding of the logic underlying a domain of mathematics, our results on the structure of this exposition, and in particular on its evolution, should approximate analogous results on the latent structure of mathematics. As such, methods dealing with the emergence of meso-scale features could prove useful in studies of learning [51]. For example, how do semantic networks of students' knowledge evolve as students are taught? Can that evolution be formally predicted by a generative network model built from the textbook used in their class? And can students' performance in a class be predicted by the extent to which their knowledge reflects the latent structure of the subject matter?

(b) Knowledge gaps in the exposition of mathematics texts

Using the tool of persistent homology, we examined the growth and persistence of knowledge gaps (more explicitly topological cavities) in the semantic networks of linear algebra texts. While this tool has been applied to other types of text and knowledge, including Shakespeare's plays [52], natural language [53], discourse [54] and collections of mathematics papers [55], little is known about how gaps within a single expositional text or growing semantic network may impact how that text or knowledge structure might be received or understood. Our hypothesis, motivated by the idea that a topologically complex structure with many gaps in knowledge might be more difficult to learn, was that effective exposition likely seeks to produce a smaller number of knowledge gaps, as the creation of a great number of topological cavities could prove confusing to a reader. Still, leaving a few gaps throughout exposition can add intrigue to the subject, piquing the reader's curiosity to make connections themselves [56].

In the context of this discussion, it is interesting to contrast the features of a process that humans have arguably optimized for explicit learning with the features of a process that nature has arguably optimized for implicit learning [57]. As a token of the former, we consider textbook writing; as a token of the latter, we consider language acquisition in children [6,58]. Evidence suggests that knowledge gaps, detected as topological cavities, are a robust feature of language acquisition in toddlers and their prevalence is unaffected by maternal education or by the order in which words are learned [28]. One could speculate that this observed homogeneity in the early semantic feature network learning supports robust language acquisition, ensuring that children who are exposed to different sets of words at different times are still able to reach adult language proficiency. By contrast, when constructing an exposition for a textbook whose sole purpose is to take a set of students from naivety to sophistication in the same place and at the same time, such robustness is not needed and instead consistency, thoroughness and comprehensiveness is required. The inability of randomized models to reliably produce cycle lifetimes similar to those seen in the textbooks we study here would be consistent with these distinctions in goals and environment. It could prove useful in the future to more generically assess the robustness of growing networks to the order of node introduction [59], particularly to assess differences between implicit or explicit learning processes.

Notably, we found that most cavities were eliminated before the end of each text. We observed that, while multiple connected components were introduced, all were eventually—and usually quite quickly—connected into a single connected component, suggesting that the expositional order of introduction of edges throughout the text minimizes the extent to which cavities are formed. Remarkably, though, the order of the expositions—that is, the extent to which cycles were not introduced and did not persist—did not appear to be maximal. That is, the node-ordered filtration null model exhibited significantly sparser persistent homology than we observed in the texts (electronic supplementary material, figure S8). This observation suggests a trade-off between topological order and apparent learnability; specifically, while such neatly ordered expositions might minimize the extent to which knowledge gaps are created and persist, it is likely in the best interest of readable and enjoyable exposition to not follow this purely structural ordering—that is, to properly motivate concepts, give relationships where they might seem natural and useful, and make the text generally more readable.

Our correlation analysis of the barcode densities suggests some interesting directions for further study of the potential relationship between persistent homology of a growing semantic network and effective learnability. Specifically, while our study did not deal explicitly with differential learnability of texts or in how knowledge gaps might affect the learning process, we did observe several interesting relationships between the 0- and 2-dimensional barcode densities and textbook ratings. While these results are preliminary, they suggest that an interesting avenue for further study would be to examine the topology of growing semantic networks in the classroom setting. In particular, one could consider multiple networks: the network of the textbook being used, providing the 'latent space' of the knowledge and the relationships between concepts; the teacher's network, as provided in class to the students through lessons; and finally, the students' networks, as they develop over time while the students learn the material. An analysis of the developmental and topological relationships between all three of these classes of semantic networks could yield interesting results in how knowledge structures are transferred from teacher and book to student, and could provide useful insight to effective structuring and expositional presentation of knowledge in a textbook format.

(c) The effectiveness of the random index null model

Throughout our study, we have used the random index model as a null to examine how the results we obtain for the texts' actual semantic networks differ from what we might expect when simply calculating the co-occurrence networks and filtrations of a set of random words in a text. Notably, while most of our results have fallen at the extreme ends of the metric distributions exhibited by the random index ensemble, in some cases, such as in 1- and 2-dimensional normalized average cycle lifetime, the empirical filtrations demonstrate values that fall within the bulk of the corresponding random index ensemble results. The perspective that the null simply gives us a weighted network and filtration computed from the co-occurrence of a random set of words might be disheartening, as this could suggest that our results, rather than reflecting the meaningful structure of semantic networks of concepts elucidated by textbooks, instead might simply reflect growing topologies that would be expected from any analogous calculation of co-occurrence of arbitrarily-chosen words or phrases within a text. However, recall that the random index sets comprise words not found within the stop word list, and therefore we might consider each random index null network as a semantic network itself. Certainly, the semantic features extracted through co-occurrence of randomly chosen words might not reflect the content which is the primary focus of the text, since the random index set may include words that are not mathematically meaningful. Even so, it is likely that some mathematical words will make their way into the random index set, and the remainder of the words in the index set, due to their not being stop words, will have some (admittedly non-mathematical) meaning. Thus, the random index set may actually be viewed as an ensemble of growing semantic networks, each of which happens to have a different node set and hence extracts semantic information about the relationships between different words. From this perspective, the explanatory power of the random index model both makes sense, and should be expected.

(d) Methodological considerations

There are certain limitations inherent in our work that should be considered for future study. First, our text extraction methodology imperfectly converted PDFs to plaintext, leaving significant textual noise and artefacts of embedded math which required subsequent automated removal, and the remnants of which prevented perfect concept extraction and sentence-level co-occurrence calculation. Because textbook PDFs are easier to access than textbook source material, we spent significant time developing our text extraction approach to account for these circumstances so that our methodology could be widely applicable. However, future work could use the LaTeX source for textbooks in order to reduce noise. Second, the problem of concept extraction is ill-posed due to the subjectivity of the notion of ‘concept’. We examined a number of supervised and unsupervised keyphrase extraction algorithms, and in addition to performing the best in comparison to our intuitive expectation for linear algebra concepts, we found RAKE to be both fast and versatile, as its unsupervised nature allows it to work on any text without requiring training on a large corpus of similar documents. However, future work will be necessary to better understand (a) how to determine how many concepts should be extracted from a text, (b) what should comprise a ‘concept’ in a semantic network, and to (c) examine hierarchically structured semantic networks to incorporate the subjectivity of concepts into the network structure, so that high-level concepts are distinguished from those which are lower-level. Third, our network and filtration construction methodology is only one of many possible methodologies; as we chose to use co-occurrence to construct the networks, they are undirected and lack edge labels detailing the nature of each relationship. Fourth, the application of a clique complex to infer knowledge gaps in a growing network is one of many choices, and it assumes that any fully-connected $(k+1)$-clique should, in fact, reflect a filled k-simplex of knowledge. However, a possible alternative could be to only add a k-simplex when such higher-order relationships are observed simultaneously, such as when three words co-occur in the same sentence. Finally, further research in a classroom setting and across multiple domains of knowledge beyond mathematics could provide insight into what types of knowledge gaps might have an effect on student learning, thus providing an answer as to how persistent homology should be computed on growing semantic networks.

(e) Future directions

An open area for future work lies in understanding trade-offs in ordered network structure. Here, we find four separate instances in which semantic networks of linear algebra textbooks appear to balance competing constraints. First, while core-ness and modularity are higher than expected in a continuous configuration null model, they are notably lower than expected in a random index null model. Second, while core nodes tend to be added more quickly than periphery nodes, the difference in speed is more stark in the random index model. Third, while some texts add core nodes faster than expected in the random sentence order null model, some texts add core nodes more slowly, suggesting that each text opts for a different expositional style. Fourth, while the barcodes of the empirical networks are relatively sparse compared to the random edge model, they still exhibit more persistent cycles than the most ordered model, the node-ordered filtration. Collectively, these results suggest that effective and useful exposition, while structured in nature, is not as strongly structured as it could be. It may be effective to purposefully introduce some gaps in knowledge by withholding topics to support productive failure [25] or provide detailed motivation to stimulate curiosity [56,60]. Of course, it is also possible that our observations reflect the nature of the structure of mathematics: perhaps mathematics simply does not have as strongly ordered a structure as we might observe in our null models. Future efforts across multiple disciplines could seek to better understand this tradeoff and its potential causes.

4. Material and methods

All tools and methods developed for use in this work are designed to be broadly applicable to any expositional text. We thus provide Python code for the extraction and analysis of semantic networks at https://github.com/nhchristianson/Math-text-semantic-networks. Further details and considerations for the methods used can be found in the electronic supplementary material, Methods.

(a) Data collection and preprocessing

We collected a diverse set of 10 linear algebra textbooks in PDF format, ranging in focus from theory to application (see electronic supplementary material, Methods for more details). We converted the PDF files to plaintext with the tool at https://pdftotext.com, and manually cleaned each text to isolate the main chapters, discarding introductory or appendix sections. We then converted the text to unicode KD normal form, replaced hyphens with spaces, and used spaCy [61] to lemmatize all words in each text, which reduces inflected words to their dictionary form. We then used the Python Natural Language Toolkit (NLTK, v. 3.3 [62]) to tokenize the text into sentences and their component words, replacing any word containing numerical characters with the character '#'. We then removed all words not comprised solely of letters and '#' and made the remaining words lowercase. Due to the presence of embedded mathematics in the textbooks which is imperfectly handled by the PDF-to-text conversion process, we implemented a final measure in an attempt to clean the text of artefacts such as the remnants of variables and equations. In particular, we first created a 'stop list'—that is, a list of common 'stop words' in English—by removing all single-letter words from the Ranks NL Long stop word list (https://www.ranks.nl/stopwords), since such single character words likely represent variables within the text. Then, we applied a series of rules to determine whether any word token was sufficiently variable-like to be converted to a 'VAR' variable placeholder: any '#' placeholder was kept as is; then, any word without vowels was converted to 'VAR'; then, any word of length at most two not present on our stop list was converted to 'VAR'; and finally, any word of length 3 or 4 that did not spell check using the Enchant [63] spell-checker was converted to 'VAR'.

(b) Concept extraction

The linguistics and natural language processing literature provide a number of canonical statistical metrics for determining the significance of *n-grams*, or phrases comprised *n* words, within text [64]. After testing multiple supervised and unsupervised keyphrase extraction methodologies, we chose to use an unsupervised method based on the rapid automatic keyword extraction (RAKE) algorithm [33] to extract concepts from our texts. RAKE works as follows:

(i) A provided set of stop words, phrase delimiters, and word delimiters are used to divide the document into a set of candidate keyphrases and their comprising keywords.
(ii) The frequency of keywords and their co-occurrence in different keyphrases is calculated, forming a co-occurrence graph.
(iii) The candidate keyphrases are ranked by a scoring function score(k), which typically ranks candidate keyphrases by certain properties of their comprising keywords.
(iv) Some threshold n is chosen, and the top n ranked candidate keyphrases are kept as the extracted keyphrases.

In RAKE, the scoring function for a candidate keyphrase k is typically taken to be

$$\mathrm{score}_{\mathrm{RAKE}}(k) = \sum_i \frac{\deg(k_i)}{\mathrm{freq}(k_i)},$$

where $\deg(k_i)$ and $\mathrm{freq}(k_i)$ are the degree and frequency, respectively, of the ith keyword comprising the phrase k in the co-occurrence graph RAKE constructs. As such, RAKE poses that significant keyphrases are those whose component words co-occur with many other words, but do not occur very frequently. Because we wish to ensure that the scores of more plausibly mathematical words are high, we modify this keyphrase scoring function to incorporate the term frequency-inverse document frequency ranking method [65], including an additional term to account for a given keyphrase's frequency in an external corpus. Specifically, we specify our phrase scoring function as

$$\mathrm{score}(k) = \frac{\mathrm{score}_{\mathrm{RAKE}}(k)}{1 + \mathrm{brown}(k)},$$

where brown(k) is the number of times that the whole keyphrase occurs in the Brown corpus [66] with the 'Learned' category (comprised of scientific and other academic texts) removed. As such, we aim to penalize phrases that occur very frequently in non-mathematical text, as such words will likely not be mathematically meaningful. We add 1 to the brown(k) term in the denominator since not all phrases RAKE extracts occur in the Brown corpus. Details on our specific implementation of the modified RAKE algorithm can be found in the electronic supplementary material, Methods.

(c) Network construction

We construct each text's semantic network of concepts by calculating the co-occurrence of concepts in each text on a sentence level. That is, we deem two concepts in each text's index set to co-occur, and thus be related, if they occur in the same sentence at some point in the text. We also assign to each edge between concepts an integer weight indicating the number of sentences in which the two concepts co-occur. These data yield an undirected weighted graph $G = (V, E)$, where each node $v \in V$ is a concept and each edge $(v_1, v_2) = e \in E$ represents a semantic relationship between concepts with an associated positive integer weight $w(e) \in \mathbb{Z}^+$ denoting the number of sentences in which the two concepts co-occur.

We are not merely interested in the total semantic network of each textbook, but in the development of the semantic networks over the course of exposition. Thus, for each text, we keep track of the first sentence in which each concept and each relationship—equivalently, each node and each edge—is introduced. If a text has N sentences, our methodology of extracting growing semantic networks yields a sequence of N graphs $G_1 \to \cdots \to G_N$, where the kth graph G_k includes all nodes and edges which have been introduced prior to or during the kth sentence of the text. In the context of algebraic topology, which we employ throughout this study, such a sequence of nested objects is called a filtration. We call this sequence of graphs the *expositional filtration* of a text. In considering this filtration, we consider the binarized graphs; that is, we disregard edge weight data during the exposition, only considering edge weight data for the final semantic network, which we call the *total network*.

(d) Meso-scale network structure

Complex networks often exhibit meso-scale or global characteristics of structural order. Certain networks exhibit *community structure*, in which densely connected *communities* of nodes exhibit sparse or weak inter-community connections [67]. In the context of semantic networks, such densely connected communities may represent strongly related concepts that indicate the existence of some higher-order enveloping concept or umbrella term. Another type of meso-scale structure which may be exhibited is *core–periphery* structure, which is characterized by a densely connected set of *core* nodes and a set of *periphery* nodes which are sparsely connected amongst themselves, but are strongly connected to the core [68]. Such an organization of semantic networks is plausible in the context of mathematics, in which many different ideas may be developed from a smaller set of highly related concepts.

To detect community and core–periphery structure in the networks, we used the Brain Connectivity Toolbox for Python, v. 0.5.0, which is based on the Matlab Brain Connectivity Toolbox (BCT) [69]. To evaluate the presence of a core–periphery structure, we seek to assign a network's nodes to either the core or the periphery group so as to maximize the *core-ness* quality function [70]:

$$Q_C = \frac{1}{v_C}\left(\sum_{i,j\in C_c}(w_{ij}-\gamma_C\bar{w}) - \sum_{i,j\in C_p}(w_{ij}-\gamma_C\bar{w})\right),$$

where C_c and C_p are the sets of nodes in the core and periphery, respectively, w_{ij} is the weight of the edge from node i to node j (which will be 0 if the nodes are not connected by an edge), $\bar{w}$ is the average of all edge weights, where non-existent edges with 'zero weight' are also included in the average, γ_C is a resolution which controls the size of the core, which we set to 1, and v_C is a normalization constant. In effect, in maximizing core-ness we seek to maximize the number and weight of intra-core connections, while minimizing the number and weight of intra-periphery connections.

To evaluate the presence of community structure in the networks, we use a Louvain-like locally greedy algorithm [71] to optimize the modularity quality function:

$$Q_M = \frac{1}{v_M}\sum_{i,j\in C}\left(w_{ij}-\gamma_M\frac{s_i s_j}{v_M}\right)\delta_{ij},$$

where C is the set of network nodes, w_{ij} is the weight of the connection from node i to node j, s_i and s_j are the summed weights of edges connected to node i and node j, respectively, γ_M is a resolution parameter controlling the size of communities which we set to 1, v_M is a normalization constant, and δ_{ij} is the Kronecker delta function, which is 1 when node i and node j are in the same community and is 0 otherwise [70]. In effect, modularity maximization seeks to maximize the strength and number of connections within communities, yielding a partition of the network nodes into a set of densely connected communities with few inter-community connections.

(e) Persistent homology

Beyond characterization of the local and meso-scale attributes of the total semantic network of the texts, we furthermore seek to evaluate structural and topological characteristics of the semantic networks as they are built over the course of the entire text. In particular, we study the extent to which 'knowledge gaps' are created and persist in semantic networks throughout a text's exposition. To this end, we use a method with roots in the mathematics of algebraic topology called *persistent homology* which, in short, evaluates the creation and lifespan of topological 'holes' in data over time, or in this case, over the course of exposition, thus allowing us to characterize and evaluate the presence of these gaps in knowledge. Here we give a brief, intuitive overview of how we calculate persistent homology for our expositional semantic networks; the particularly interested reader may refer to [29–31] for a rigorous overview of persistent homology and its computation for data analysis, as well as [32,72–75] for example uses of persistent homology in the context of complex networks.

Recall that a text's semantic network at a certain point in the exposition (a particular graph in the expositional filtration) is an undirected graph, where connections between nodes indicate that the concepts represented by those nodes have already co-occurred in a sentence. Given a binary undirected graph $G = (V, E)$, we may construct an object called the *clique complex* $X(G)$, which, for every natural number k, assigns to every all-to-all connected subgraph of G on $(k + 1)$ vertices (also known as a $(k + 1)$-clique) a k-*simplex*, which may geometrically be represented as the convex hull of $(k + 1)$ affinely independent points. For example, a 0-simplex is simply a single node, a 1-simplex is an edge, a 2-simplex is a filled-in triangle and a 3-simplex is a filled-in tetrahedron. Intuitively speaking, this clique complex $X(G)$ is a 'filled-in' version of the graph G, where, for each k, we choose a distinct colour and then colour in all $(k + 1)$-cliques in G to form k-simplices. Then, classical *homology* intuitively describes, for each k, how many topological 'holes' of dimension k are in the complex, or how many regions are enclosed by the kth colour, but are themselves not coloured. In other words, homology detects *cycles*[1] of k-simplices that surround a void. For example, a 1-cycle reflects a conventional cycle in a graph, just like the hole in a circle; and a 2-cycle reflects a cavity, like the hole in the centre of a sphere. A 0-cycle is intuitively slightly different, in that 0-cycles refer to connected components of the graph, so that having more than one 0-cycle tells us that multiple disconnected components exist. In our work, we restrict our focus to these first three dimensions, since these are the most geometrically intuitive. These cavities or holes are exactly the knowledge gaps we seek in the semantic networks, as they indicate some closed cycle of $(k + 1)$-order connections between concepts surrounding a region of lesser connectivity.

A useful extension of homology enables us both to count the number of holes present in the semantic network at each step in a text's exposition, as well as to keep track of which topological cavities are created and destroyed at each step. Specifically, *persistent homology* allows for the computation of the homology for the sequence of clique complexes of our expositional filtration $X(G_1) \to \cdots \to X(G_N)$; this tool not only keeps track of the number of cavities of each dimension present at each expositional step, but it also tracks the *persistence* of each individual cavity over the course of the exposition, so we may identify individual knowledge gaps, when they were created, how long they persist, and when they are extinguished. Rigorously, the kth persistent homology of a graph filtration yields a (multi-)set of intervals called the *barcode*:

$$\{[b_1, d_1), \ldots, [b_m, d_m)\},$$

where b_i indicates the time of birth of the ith k-dimensional cavity, and d_i indicates the time of death of that cavity (which may be ∞ if the cavity is still present in the total network, i.e. it never dies). Thus, the number of intervals, as well as their length (the difference between their death and birth times), indicate the number and persistence, respectively, of topological cavities during exposition.

Once we have computed the persistent homology of a text's expositional filtration for a given k, we may use several characteristics of the resultant persistence intervals to examine various aspects of the persistence of knowledge gaps in the semantic networks. In particular, we consider two metrics: first, we examine the value of m, which gives the total number of k-cavities which were created, at some point, over the course of the exposition. Secondly, we define a metric similar to one presented in [36], which we refer to as the *normalized average cycle lifetime* of dimension k:

$$D_k = \frac{1}{mN} \sum_{i=1}^{m} (d_i - b_i),$$

[1] More precisely, homology finds equivalence classes of cycles, but we refer to an equivalence class as a cycle for simplicity.

where N is the number of steps in the filtration, and d_i is the time of death of the ith k-cavity, unless $d_i = \infty$, in which case we set $d_i = N + 1$, to distinguish these infinitely persisting cavities from those that die at step N. Intuitively, this metric describes the extent to which an expositional filtration has cycles which are persistent; it is normalized by N, the length of the filtration, and m, the number of k-cycles introduced throughout the filtration, so as to be comparable across texts which might have different filtration lengths or total numbers of cycles introduced. The goal is to allow a formal comparison of how persistent k-cycles tend to be in different texts.

In our work, we use Ripser.py [76] due to its speed and efficiency for the computation of persistent homology for the empirical networks and the null models.

(f) Null models

In order to determine to what extent the results we obtain for meso-scale structure and topological dynamics in the semantic networks are significant, we employ two categories of null models: data-level null models, which randomize on the scale of the underlying text and index list, from which we may then extract semantic networks and expositional filtrations; and projected network-level null models, which randomize on the scale of the networks we extract for each text. Furthermore, while some of these models are particularly suited as null models for the structural metrics on the total network since they yield a single, weighted network, others are more suitable as null models for the growing dynamics of the semantic networks, as they provide a null expositional filtration. For each null model, our null ensemble is comprised of 100 random instantiations; we present the resulting null distributions of metrics alongside the data for our actual networks in our results. Here we summarize the null models and their uses; more detailed descriptions of the models can be found in the electronic supplementary material, Methods.

(a) Random index: Expositional filtration of words that are not common ‘stop words’ (see electronic supplementary material, Methods) chosen at random from each text, bypassing the RAKE keyphrase extraction methodology, to serve as a semantic network on random ‘concepts’. Acts as a null model for the total network and empirical filtration.

(b) Random sentence order: Expositional filtration of original index set with randomly shuffled sentences, keeping sub-sentence structure while randomizing exposition on the sentence scale. Acts as a null model for the empirical filtration.

(c) Continuous configuration: Rewired network preserving node degree and strength. Acts as a null model for the total network.

(d) Random edge: Random reordering of edge introduction from the empirical filtration, mimicking totally random exposition. Acts as a null model for the empirical filtration.

(e) Node-ordered: Adds each node and all its edges in order of node introduction in the exposition, mimicking exposition that connects each concept to all previously-related concepts. Acts as a null model for the empirical filtration.

Data accessibility. All code used for the extraction and analysis of semantic networks in this work is accessible at https://github.com/nhchristianson/Math-text-semantic-networks.

Authors’ contributions. N.H.C. defined the research strategy. N.H.C. and D.S.B. secured funding. N.H.C. performed the analysis and led the project. N.H.C. wrote the paper. A.S.B. and D.S.B. edited the paper.

Competing interests. We declare we have no competing interests

Funding. D.S.B., N.H.C., and A.S.B. acknowledge support from the John D. and Catherine T. MacArthur Foundation, the Alfred P. Sloan Foundation, the ISI Foundation, the Paul Allen Foundation, the Army Research Laboratory (W911NF-10-2-0022), the Army Research Office (Bassett-W911NF-14-1-0679, Grafton-W911NF-16-1-0474, DCIST- W911NF-17-2-0181), the Office of Naval Research, the National Institute of Mental Health (2-R01-DC-009209-11, R01-MH112847, R01-MH107235, R21-M MH-106799), the National Institute of Child Health and Human Development (1R01HD086888-01), National Institute of Neurological Disorders and Stroke (R01 NS099348) and the National Science Foundation (BCS-1441502, BCS-1430087, NSF PHY-1554488 and BCS-1631550).

Acknowledgements. We are grateful to David Lydon-Staley, Dale Zhou, Alec Helm and Shubhankar Patankar for their generous advice on early versions of this manuscript.

References

1. Sowa JF. 2006 Semantic networks. In *Encyclopedia of cognitive science* (ed. L Nadel). New York, NY: John Wiley & Sons.
2. Steup M. 2018 Epistemology. In *The Stanford encyclopedia of philosophy* (ed. EN Zalta). Metaphysics Research Lab, Stanford University Winter 2018 edition.
3. Hartley RT, Barnden JA. 1997 Semantic networks: visualizations of knowledge. *Trends Cogn. Sci.* **1**, 169–175. (doi:10.1016/S1364-6613(97)01057-7)
4. Lehmann F. 1992 Semantic networks. *Comput. Math. with Appl.* **23**, 1–50. (doi:10.1016/0898-1221(92)90135-5)
5. Nickel M, Murphy K, Tresp V, Gabrilovich E. 2016 A review of relational machine learning for knowledge graphs. *Proc. IEEE* **104**, 11–33. (doi:10.1109/JPROC.2015.2483592)
6. Steyvers M, Tenenbaum JB. 2005 The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cogn. Sci.* **29**, 41–78. (doi:10.1207/s15516709cog2901_3)
7. Gallenstein NL. 2011 Mathematics concept maps: assessing connections. *Teaching Children Math.* **17**, 436–440.
8. Broggy J, McClelland G. 2009 Integrating concept mapping into higher education: a case study with physics education students in an Irish university. In *British Education Research Association Annual Conference, Manchester University, 2–5 September*.
9. Hill LH. 2005 Concept mapping to encourage meaningful student learning. *Adult Learning* **16**, 7–13. (doi:10.1177/104515950501600302)
10. Daley BJ. 2002 Facilitating learning with adult students through concept mapping. *J. Continuing Higher Edu.* **50**, 21–31. (doi:10.1080/07377366.2002.10401192)
11. Lapp DA, Nyman MA, Berry JS. 2010 Student connections of linear algebra concepts: an analysis of concept maps. *Int. J. Math. Edu. Sci. Technol.* **41**, 1–18. (doi:10.1080/00207390903236665)
12. Newman MEJ. 2010 *Networks: an introduction*. Oxford, UK: Oxford University Press.
13. Utsumi A. 2015 A complex network approach to distributional semantic models. *PLoS ONE* **10**, e0136277. (doi:10.1371/journal.pone.0136277)
14. Lynn CW, Papadopoulos L, Kahn AE, Bassett DS. 2020 Human information processing in complex networks. *Nat. Phys.* 1–9. (doi:10.1038/s41567-020-0924-7)
15. Lynn CW, Kahn AE, Nyema N, Bassett DS. 2020 Abstract representations of events arise from mental errors in learning and memory. *Nat. Commun.* **11**, 2313. (doi:10.1038/s41467-020-15146-7)
16. Kahn AE, Karuza EA, Vettel JM, Bassett DS. 2018 Network constraints on learnability of probabilistic motor sequences. *Nat. Hum. Behav.* **2**, 936–947. (doi:10.1038/s41562-018-0463-8)
17. da Fontoura Costa L. 2006 Learning about knowledge: a complex network approach. *Phys. Rev. E* **74**, 026103. (doi:10.1103/PhysRevE.74.026103)
18. Karuza E, Thompson-Schill S, Bassett D. 2016 Local patterns to global architectures: influences of network topology on human learning. *Trends Cogn. Sci.* **20**, 629–640. (doi:10.1016/j.tics.2016.06.003)
19. Koponen IT, Nousiainen M. 2018 Concept networks of students' knowledge of relationships between physics concepts: finding key concepts and their epistemic support. *Appl. Netw. Sci.* **3**, 14. (doi:10.1007/s41109-018-0072-5)
20. Yun E, Park Y. 2018 Extraction of scientific semantic networks from science textbooks and comparison with science teachers' spoken language by text network analysis. *Int. J. Sci. Edu.* **40**, 2118–2136. (doi:10.1080/09500693.2018.1521536)
21. Fang Z, Wang J, Liu B, Gong W. 2011 Wikipedia as domain knowledge networks - domain extraction and statistical measurement. In *International Conference on Knowledge Discovery and Information Retrieval (KDIR) 2011, Paris, France, 26–29 October* (eds J Filipe, A Fred). SciTePress.
22. Chai LR, Zhou D, Bassett DS. 2020 Evolution of semantic networks in biomedical texts. *J. Complex Netw.* **8**, cnz023. (doi:10.1093/comnet/cnz023)
23. Pereira H, Fadigas I, Senna V, Moret M. 2011 Semantic networks based on titles of scientific papers. *Physica A* **390**, 1192–1197. (doi:10.1016/j.physa.2010.12.001)

24. Ritter F, Nerb J, Lehtinen E, O'Shea T (eds). 2007 *In order to learn: how the sequence of topics influences learning*. Oxford, UK: Oxford University Press.
25. Kapur M. 2014 Productive failure in learning math. *Cogn. Sci.* **38**, 1008–1022. (doi:10.1111/cogs.12107)
26. Mowat E. 2008 Making Connections: mathematical understanding and network theory. *Learning Math.* **28**, 20–27.
27. Mowat E, Davis B. 2010 Interpreting embodied mathematics using network theory: implications for mathematics education. *Complicity Inte. J. Complex. Educ.* **7**, 1–31. (doi:10.29173/cmplct8834)
28. Sizemore AE, Karuza E, Giusti C, Bassett D. 2019 Knowledge gaps in the early growth of semantic feature networks. *Nat. Hum. Behav.* **2**, 682–692. (doi:10.1038/s41562-018-0422-4)
29. Carlsson G. 2009 Topology and data. *Bull. Amer. Math. Soc.* **46**, 255–308. (doi:10.1090/S0273-0979-09-01249-X)
30. Zomorodian A, Carlsson G. 2005 Computing persistent homology. *Discrete Comput. Geom.* **33**, 249–274. (doi:10.1007/s00454-004-1146-y)
31. Edelsbrunner H, Morozov D. 2013 Persistent homology: theory and practice. In *Proc. of the European Congress of Mathematics*, pp. 31–50. (doi:10.4171/120-1/3)
32. Otter N, Porter MA, Tillmann U, Grindrod P, Harrington HA. 2017 A roadmap for the computation of persistent homology. *EPJ Data Sci.* **6**, 17. (doi:10.1140/epjds/s13688-017-0109-5)
33. Rose S, Engel D, Cramer N, Cowley W. 2010 Automatic keyword extraction from individual documents. In *Text mining: applications and theory* (eds M Berry, J Kogan), pp. 1–20. New York, NY: John Wiley & Sons, Ltd.
34. Peixoto TP. 2014 The graph-tool python library. See http://figshare.com/articles/graph_tool/1164194.
35. Virtanen P *et al.* 2020 SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272. (doi:10.1038/s41592-019-0686-2)
36. Adcock A, Carlsson E, Carlsson G. 2016 The ring of algebraic functions on persistence bar codes. *Homol. Homotopy Appl.* **18**, 381–402. (doi:10.4310/HHA.2016.v18.n1.a21)
37. Hart K. 1981 Hierarchies in mathematics education. *Edu. Stud. Math.* **12**, 205–218. (doi:10.1007/BF00305622)
38. Viteri S, DeDeo S. 2020 Explosive proofs of mathematical truths. (http://arxiv.org/abs/2004.00055)
39. Gherardi M, Rotondo P. 2016 Measuring logic complexity can guide pattern discovery in empirical systems. *Complexity* **21**, 397–408. (doi:10.1002/cplx.21819)
40. Gijselaers WH. 1996 Connecting problem-based practices with educational theory. *New Dir. Teaching Learn.* **1996**, 13–21. (doi:10.1002/tl.37219966805)
41. Avital SM. 1973 Teaching a mathematical proof by exposition or the 'Let us Define a Function' syndrome. *Int. J. Math. Edu. Sci. Technol.* **4**, 143–147. (doi:10.1080/0020739730040209)
42. Fried MN, Jahnke HN. 2015 Otto Toeplitz's 1927 Paper on the genetic method in the teaching of mathematics. *Sci. Context* **28**, 285–295. (doi:10.1017/S0269889715000034)
43. Börner K, Chen C, Boyack KW. 2003 Visualizing knowledge domains. *Annu. Rev. Inf. Sci. Technol.* **37**, 179–255. (doi:10.1002/aris.1440370106)
44. Chen B, Tsutsui S, Ding Y, Ma F. 2017 Understanding the topic evolution in a scientific domain: an exploratory study for the field of information retrieval. *J. Informetrics* **11**, 1175–1189. (doi:10.1016/j.joi.2017.10.003)
45. Dworkin JD, Shinohara RT, Bassett DS. 2018 The landscape of NeuroImage-ing research. *NeuroImage* **183**, 872–883. (doi:10.1016/j.neuroimage.2018.09.005)
46. Dworkin JD, Shinohara RT, Bassett DS. 2019 The emergent integrated network structure of scientific research. *PLoS ONE* **14**, e0216146. (doi:10.1371/journal.pone.0216146)
47. Csermely P, London A, Wu LY, Uzzi B. 2013 Structure and dynamics of core/periphery networks. *J. Complex Netw.* **1**, 93–123. (doi:10.1093/comnet/cnt016)
48. Verma T, Russmann F, Araujo NAM, Nagler J, Herrmann HJ. 2016 Emergence of core-peripheries in networks. *Nat. Commun.* **7**, article 10441. (doi:10.1038/ncomms10441)
49. Bassett DS, Porter MA, Wymbs NF, Grafton ST, Carlson JM, Mucha PJ. 2013 Robust detection of dynamic community structure in networks. *Chaos* **23**, 013142. (doi:10.1063/1.4790830)
50. Alvari H, Hajibagheri A, Sukthankar G, Lakkaraju K. 2016 Identifying community structures in dynamic networks. *Soc. Netw. Anal. Min.* **6**, 77. (doi:10.1007/s13278-016-0390-5)

51. Zurn P, Bassett DS. 2020 Network architectures supporting learnability. *Phil. Trans. R. Soc. B* **375**, 20190323. (doi:10.1098/rstb.2019.0323)
52. Rieck B, Leitte H. 2016 ‘Shall I compare thee to a network?’: Visualizing the Topological Structure of Shakespeare’s plays. In *Workshop on Visualization for the Digital Humanities at IEEE VIS, Baltimore, MD, 23–28 October*. Piscataway, NJ: IEEE.
53. Zhu X. 2013 Persistent homology: an introduction and a new text representation for natural language processing. In *Proc. of the 23rd Int. Joint Conf. on Artificial Intelligence, IJCAI ’13, Beijing, China, 3–9 August*, pp. 1953–1959. AAAI Press.
54. Savle K, Zadrozny W, Lee M. 2019 Topological data analysis for discourse semantics?. In *Proc. of the 13th Int. Conf. on Computational Semantics - Student Papers*, pp. 34–43 Gothenburg, Sweden. Association for Computational Linguistics.
55. Salnikov V, Cassese D, Lambiotte R, Jones NS. 2018 Co-occurrence simplicial complexes in mathematics: identifying the holes of knowledge. *Appl. Netw. Sci.* **3**, 37. (doi:10.1007/s41109-018-0074-3)
56. Loewenstein G. 1994 The psychology of curiosity: a review and reinterpretation. *Psychol. Bull.* **116**, 75–98. (doi:10.1037/0033-2909.116.1.75)
57. Seger CA. 1994 Implicit learning. *Psychol. Bull.* **115**, 163–196. (doi:10.1037/0033-2909.115.2.163)
58. Hills TT, Maouene M, Maouene J, Sheya A, Smith L. 2009 Longitudinal analysis of early semantic networks preferential attachment or preferential acquisition? *Psychol. Sci.* **20**, 729–739. (doi:10.1111/j.1467-9280.2009.02365.x)
59. Sizemore Blevins A, Bassett DS. 2020 Reorderability of node-filtered order complexes. *Phys. Rev. E* **101**, 052311. (doi:10.1103/PhysRevE.101.052311)
60. Wade S, Kidd C. 2019 The role of prior knowledge and curiosity in learning. *Psychon. Bull. Rev.* **26**, 1377–1387. (doi:10.3758/s13423-019-01598-6)
61. Honnibal M, Montani I. 2017 spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *Convolut. Neural Networks Incremental Parsing*. To appear.
62. Bird S, Klein E, Loper E. 2009 *Natural language processing with Python*. Sebastopol, CA: O’Reilly Media.
63. Lachowicz D. 2017 Enchant. See https://www.abisource.com/projects/enchant/.
64. Manning CD, Schütze H. 1999 Collocations. In *Foundations of statistical natural language processing*. New York, NY: The MIT Press.
65. Salton G, Buckley C. 1988 Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.* **24**, 513–523. (doi:10.1016/0306-4573(88)90021-0)
66. Kučera H. 1967 *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
67. Newman MEJ. 2006 Modularity and community structure in networks. *Proc. Natl Acad. Sci. USA* **103**, 8577–8582. (doi:10.1073/pnas.0601602103)
68. Borgatti SP, Everett MG. 2000 Models of core/periphery structures. *Soc. Netw.* **21**, 375–395. (doi:10.1016/S0378-8733(99)00019-2)
69. Rubinov M, Sporns O. 2010 Complex network measures of brain connectivity: uses and interpretations. *NeuroImage* **52**, 1059–1069. (doi:10.1016/j.neuroimage.2009.10.003)
70. Rubinov M, Ypma RJF, Watson C, Bullmore ET. 2015 Wiring cost and topological participation of the mouse brain connectome. *Proc. Natl Acad. Sci. USA* **112**, 10 032–10 037. (doi:10.1073/pnas.1420315112)
71. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. 2008 Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.* **2008**, P10008. (doi:10.1088/1742-5468/2008/10/P10008)
72. Horak D, Maletić S, Rajković M. 2009 Persistent homology of complex networks. *J. Stat. Mech: Theory Exp.* **2009**, P03034. (doi:10.1088/1742-5468/2009/03/P03034)
73. Petri G, Scolamiero M, Donato I, Vaccarino F. 2013 Topological strata of weighted complex networks. *PLoS ONE* **8**, e66506. (doi:10.1371/journal.pone.0066506)
74. Stolz BJ, Harrington HA, Porter MA. 2017 Persistent homology of time-dependent functional networks constructed from coupled time series. *Chaos* **27**, 047410. (doi:10.1063/1.4978997)
75. Bampasidou M, Gentimis T. 2014 Modeling collaborations with persistent homology. (http://arxiv.org/abs/1403.5346)
76. Tralie C, Saul N, Bar-On R. 2018 Ripser.py: a lean persistent homology library for Python. *J. Open Source Softw.* **3**, 925. (doi:10.21105/joss.00925)