Persistent Homology of Topic Networks for the Prediction of Reader Curiosity

Manuel D. S. Hopp, Vincent Labatut, Arthur Amalvy, Richard Dufour, Hannah Stone, Hayley Jach, Kou Murayama

Persistent Homology of Topic Networks for the Prediction of Reader Curiosity

Manuel D. S. Hopp[1], Vincent Labatut[2], Arthur Amalvy[2], Richard Dufour[3], Hannah Stone[4], Hayley Jach[5,1], Kou Murayama[1]

[1]Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Germany
{firstname.lastname}@uni-tuebingen.de,
[2]Laboratoire Informatique d'Avignon – LIA UPR 4128, Avignon Université, France
{firstname.lastname}@univ-avignon.fr,
[3]Laboratoire des Sciences du Numérique de Nantes – LS2N UMR 6004, Nantes Université, France
richard.dufour@univ-nantes.fr,
[4]Emerald Publishing, Reading, United Kingdom
hannah.j.stone@reading.ac.uk,
[5]Melbourne School of Psychological Sciences, The University of Melbourne, Australia
hayley.jach@unimelb.edu.au

Abstract

Reader curiosity, the drive to seek information, is crucial for textual engagement, yet remains relatively underexplored in NLP. Building on Loewenstein's Information Gap Theory, we introduce a framework that models reader curiosity by quantifying semantic information gaps within a text's semantic structure. Our approach leverages BERTopic-inspired topic modeling and persistent homology to analyze the evolving topology (connected components, cycles, voids) of a dynamic semantic network derived from text segments, treating these features as proxies for information gaps. To empirically evaluate this pipeline, we collect reader curiosity ratings from participants ($n = 49$) as they read S. Collins's "The Hunger Games" novel. We then use the topological features from our pipeline as independent variables to predict these ratings, and experimentally show that they significantly improve curiosity prediction compared to a baseline model (73% vs. 30% explained deviance), validating our approach. This pipeline offers a new computational method for analyzing text structure and its relation to reader engagement.

1 Introduction

Reader curiosity refers to the cognitive and affective drive that motivates individuals to seek additional information while reading. In the context of textual engagement, this can manifest as a reader's urge to continue reading, explore related topics, or seek clarifications (Schiefele, 1999). While curiosity is often studied in educational psychology, its computational modeling in natural language processing (NLP) remains relatively underexplored. Existing approaches to modeling reader engagement often rely on linguistic features, e.g., sentiment analysis, readability scores (Sotirakou et al., 2021), or word-level analyses (Berger et al., 2023; Maslej et al., 2021; Dvir et al., 2023). These methods, while valuable, primarily focus on surface-level characteristics and often fail to capture the broader semantic structure, narrative flow, and, crucially, the information gaps that stimulate curiosity. Related work leveraging knowledge graphs, while providing a richer semantic representation, does not explicitly model the reader's evolving understanding and points of information need (Abu-Salih and Alotaibi, 2024).

Building on Loewenstein's *Information Gap Theory* (Loewenstein, 1994)—where curiosity stems from recognizing a difference between known and desired information—we introduce a framework that models reader curiosity by quantifying semantic information gaps within a text's semantic structure. Unlike prior work focused on micro-level textual features, our approach adopts a macro-level, cognitive perspective, operationalizing the concept of *plot holes*, or *information gaps*, to predict reader engagement. We hypothesize that these gaps, representing areas of missing connections or coherence in the textual flow, act as intrinsic motivators. Furthermore, we operationalize surprise, a key driver of curiosity, by measuring the dynamic shifts and transformations in these information gaps throughout the text.

To realize this framework, we introduce a pipeline leveraging Topological Data Analysis

(TDA). This pipeline integrates recent topic modeling (building upon BERTopic (Grootendorst, 2022)) to extract key topics, constructs a dynamic topic network representing the flow of these topics (Zhu, 2013), and applies TDA, specifically Persistent Homology (Munch, 2017), to identify and quantify topological cavities within this network (Patankar et al., 2023; Zhou et al., 2024). These cavities—disconnected components, cycles, and voids—represent information gaps. We further employ Wasserstein and Bottleneck distances to measure the evolution of these gaps, capturing the element of surprise. This study explores the potential of this novel approach of characterizing topological features of a text.

To demonstrate feasibility, we conducted a pilot study as a proof-of-concept using S. Collins's "The Hunger Games" (Collins, 2011) novel. Participants naïve to both the book and its movie adaptation provided chapter-wise curiosity ratings, enabling an initial analysis of curiosity dynamics in response to information gaps.

Our main contributions are threefold:

1. **Pipeline:** We designed a pipeline for the modeling of textual information gaps, integrating topic modeling and TDA.

2. **Engagement Data**: We conducted a survey in order to obtain reader engagement data for "The Hunger Game" novel.

3. **Experimental validation**: We leveraged the survey data to evaluate our approach empirically.

4. **Interdisciplinary approach:** our method connects topic modeling and TDA with theories from motivational psychology, facilitating further interdisciplinary research.

We share our source code and data at `https://github.com/mds-hopp/pers_homol_data`.

The rest of this article is organized as follows. In Section 2, we review the literature directly related to our work. In Section 3, we present our narrative modeling pipeline. Section 4 describes our experimental setup, while our results are presented and discussed in Section 5. Finally, we review the salient points of our work and its perspectives in Section 6.

2 Related Work

This work builds upon and addresses limitations in computational reader engagement models, graph-based and topological methods in NLP, and network-based approaches to curiosity and exploration.

Computational Models of Reader Engagement Predicting user engagement is an important NLP task, often approached through text feature analysis. There are many features that are related to reader engagement. Early methods relied on surface characteristics like sentiment and readability (Sotirakou et al., 2021). Other approaches often relied on classical bag-of-words representations for word-level analysis (Maslej et al., 2021; Dvir et al., 2023). However, these methods, while useful, largely neglect semantic structure and cognitive processes. Other work incorporates cognitive aspects, highlighting uncertainty (Berger et al., 2023) and semantic cohesion (Ward and Litman, 2008). Yet, these typically operate at the word or sentence level, failing to model the reader's *evolving* information state –critical to theories like Loewenstein's Information Gap Theory (Loewenstein, 1994). Our pipeline directly addresses this, modeling the *dynamic* evolution of semantic information gaps. The use of text embeddings from multilingual large language models also opens the door to potential cross-cultural work on the topic, e.g. (Zhou et al., 2024).

Graph & Topological Methods in NLP The drive for higher-level text understanding has increased the use of graph representations in NLP. Knowledge graphs enhance tasks like question answering (Abu-Salih and Alotaibi, 2024), and graph-based retrieval augmented generation (RAG) methods, such as GraphRAG (Han et al., 2025) and LightRAG (Guo et al., 2024), leverage relational structure. Topic modeling also benefits from graph approaches. Traditional methods like Latent Dirichlet Allocation (LDA) are complemented by approaches using bipartite networks and community detection (Gerlach et al., 2018), and embedding-based methods like BERTopic (Grootendorst, 2022) or a combination of both embeddings and networks (Cao and Fairbanks, 2019), offering richer semantic representations. However, these generally do not analyze the *topological* structure of the resulting networks.

Topological Data Analysis (TDA), particularly

Persistent Homology, provides tools to analyze data shape, including networks. A recent review shows TDA's growing interest in NLP (Uchendu and Le, 2024). Christianson et al. (2020) used TDA to identify knowledge gaps in math textbooks, and Tymochko et al. (2021) to capture logical holes in abstracts. Critically, existing NLP applications of TDA primarily focus on *static* text representations. Our work significantly extends this, applying persistent homology to a *dynamic* topic network (inspired by the work of Zhu (2013)), tracking the evolution of topological features (specifically, cavities) over time. This dynamic aspect is crucial for modeling changing reader information gaps.

Engagement and Graph & Topological Methods Beyond NLP, network science has modeled text structure to understand its implications to learning and cognition, including curiosity-driven exploration (Zhou et al., 2024; Patankar et al., 2023). Patankar et al. (2023) tracked structural changes in time-varying graphs using persistent homology, conceptually related to our approach. However, their focus was on general graph dynamics, not reader engagement. We bridge this gap, connecting network models of exploration with cognitive theories of curiosity, specifically the Information Gap Theory, applying them to model reader engagement in NLP. Operationalizing information gaps as topological cavities in a dynamically evolving semantic network provides a novel, quantifiable measure of reader curiosity, directly linking computational methods and psychological theories. Moreover, our methodology targets continuous sequential texts (like narratives) to analyze intrinsic semantic evolution and its link to reader curiosity. To analyze how semantic evolution in continuous narratives relates to reader curiosity, we required paired text-curiosity annotations. As our literature review confirmed no dataset appropriate for tracking narrative semantic shifts exists, we collected our own for this study.

3 Methods for Narrative Modeling

We employ a combination of NLP, network analysis, and TDA to model narrative structure and its evolution, building on and extending previous work. Our analysis proceeds in three key stages:

1. **Preprocessing**. We obtain textual data and segment it using a sliding window approach.
2. **Dynamic Topic Modeling**. A dynamic topic network is built to represent the evolving thematic structure of the text. Vertices represent topics, and edges connect topics appearing in consecutive text segments, with edge weights reflecting semantic similarity.
3. **Topological Feature Extraction via Persistent Homology**. Persistent homology quantifies the network's evolving topological features representing information gaps of the topic network.

An overview of our suggested pipeline (from text preprocessing to the topological feature extraction) is depicted in Figure 1. Details regarding the data and models we employed are located in Section 4.

3.1 Dataset and Preprocessing

Our approach works with either textual or multimodal data (e.g., video), adaptable to the chosen embedding model. After data collection, the next step involves cleaning and segmenting the textual or multimodal data into smaller units. A key aspect is the collection of user engagement ratings periodically throughout content consumption. The human data allows us to validate the persistent homology measures as reader engagement.

3.2 Dynamic Topic Network

To capture the evolving thematic structure of the used texts, we construct a dynamic topic network. Each vertex in this network represents a topic and edges connect topics that occur across consecutive chunks, reflecting the narrative flow of the story. Each edge has a weight corresponding to the cosine similarity of the dyadic topics in the text embedding space.

3.2.1 Topic Modeling

We employ a pipeline inspired by the BERTopic approach for topic modeling (Grootendorst, 2022). It consists of the following three stages:

1. **Text Embedding**: Text segments are embedded into a high-dimensional vector space using the transformer-based embedding model voyage-3-large (Voyage AI, 2025). These embeddings capture the semantic meaning of the chunks, with similar chunks having closer embeddings.
2. **Dimensionality Reduction**: As a preprocessing step for clustering, we reduce the high-dimensional embedding vectors using

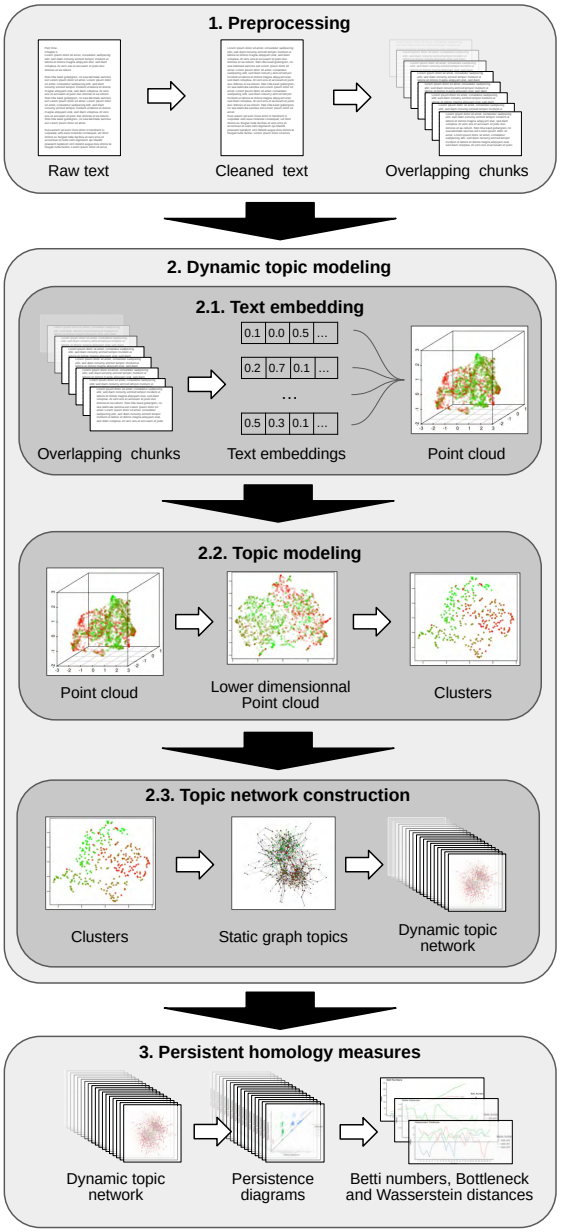

Figure 1: Pipeline from preprocessing (top) to persistent homology measures (bottom).

UMAP (Uniform Manifold Approximation and Projection, (McInnes et al., 2020)) following the BERTopic approach (Grootendorst, 2022) for improving cluster quality with high-dimensional data (Asyaky and Mandala, 2021; Allaoui et al., 2020).

3. **Clustering**: We use HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) (Campello et al., 2013), which identifies non-convex clusters of varying densities and handles noise explicitly. Furthermore, HDBSCAN automatically determines the cluster count, avoiding manual parameter tuning and potential bias. Its strong empirical performance further justifies its use (scikit-learn developers, 2025; Asyaky and Mandala, 2021; McInnes et al., 2016). Each topic is represented by the weighted centroid of its constituent text chunk embeddings, using the cluster probability as the weight.

3.2.2 Dynamic Topic Network Construction

The dynamic topic network is built incrementally, based on the measurement points of the user engagement sampling.

- **Vertices:** Each vertex in the network represents one topic.
- **Edges:** Undirected edges are created between topic vertices appearing in consecutive text chunks. This captures the sequential flow of topics throughout the narrative, as suggested by Zhu (2013) and applied in studies such as (Patankar et al., 2023).
- **Edge Weights:** The weight of each edge is determined by the cosine similarity between the embedding vectors of the connected topics in the original embedding space. Higher cosine similarity results in a stronger connection.
- **Network Series:** We segment the narrative based on user engagement rating points. Each such segment is represented by a distinct static topic graph built upon the topics and relationships occurring in the narrative up to this point. This sequence of static graphs forms a cumulative dynamic network: The first graph represents only the topics and relationships up to the first user engagement rating point, the second graph contains these topics and relations plus those occurring in the second segment, and so on. The last graph in the sequence represents all topics and relationships.

3.3 Persistent Homology Measures

To evaluate the gaps in information flow within our dynamic topic networks, we employ persistent homology, a method derived from topological data analysis. In essence, persistent homology allows us to detect and monitor the evolution of specific topological features within a network (in our case, the topic network using cosine similarity-derived distances) –namely, connected components, cycles

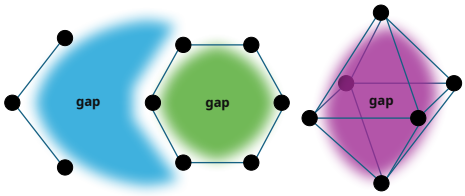

Figure 2: Identified topological features of the topic network: Components, cycles/ loops, and voids. Adapted from (Patankar et al., 2023).

(loops), and voids (enclosed empty spaces) (Moroni and Pascali, 2021; Munch, 2017)– that we compare in Figure 2.

The core of persistent homology lies in examining how a network's structure evolves as a function of a filtration parameter, denoted by ϵ. Intuitively, ϵ represents a proximity threshold. As ϵ increases, connections (edges, and in more general cases, higher-dimensional counterparts called simplices) are progressively added between vertices that are closer than the given ϵ. This process generates a nested sequence of simplicial complexes, each representing the network's structure at a specific proximity level. From this sequence, we can count the number of topological features: β_0 represents the number of connected components, β_1 the number of cycles (or loops), and β_2 the number of voids (enclosed empty spaces). Persistent homology tracks the "birth" (emergence) and "death" (merging or disappearance) of these features as ϵ increases. The results are summarized in a persistence diagram, which plots each topological feature as a point (b, d), where b signifies the ϵ value at which the feature is born, and d represents the ϵ value at which it dies. Figure 3 shows the persistence diagram for the first chapter of our dataset, revealing one highly persistent feature in dimension 1 (loop).

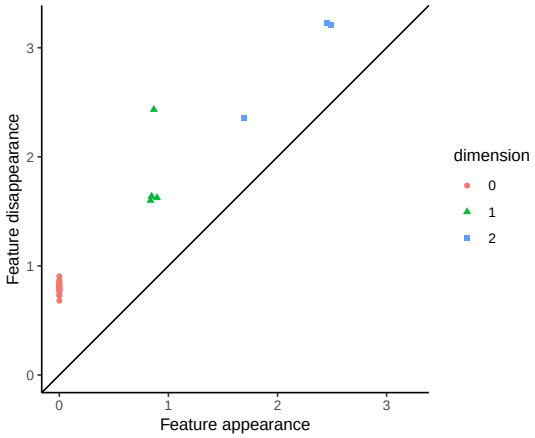

Figure 3: Persistence diagram of the first chapter of our dataset.

4 Experimental Setup

4.1 Dataset and Preprocessing

4.1.1 Dataset

The primary dataset for this pilot study consists of the young adult novel "The Hunger Games" by Collins (2011) and corresponding reader engagement data ($n = 76$ participants; mean age = 41.4, 56% female, 100% UK residence) collected through the online platform Prolific[1]. Participants were compensated in line with Prolific's recommended minimum 7.50 British Pounds per hour. Prior to reading, participants indicated their familiarity with the book and its movie adaptation. During reading, participants provided continuous self-reported ratings (0–100) on multiple engagement dimensions after each chapter. For this analysis, we focus specifically on curiosity ratings ("I was curious about this chapter") from readers unfamiliar with either the movie or book ($n = 49$). As we wished to assess the general pattern of curiosity responses (i.e., not individual reader differences, which we could not examine in this work), we computed the mean curiosity rating across participants for each chapter. Inter-Rater Reliability for mean chapter curiosity, calculated via mixed-effects models (Shrout and Fleiss, 1979) (cf. Appendix A), is .71 (moderate), showing reasonable consistency across chapters.

4.1.2 Preprocessing

We apply a two-step text preprocessing. First, we remove part titles (e.g., "Part One"), chapter titles (e.g., "The Tributes"), chapter numbers (e.g., "Chapter 1"), and empty lines. Second, we employ a sliding window approach to create text segments, using windows of 5 sentences with a 2-sentence overlap. This results in 2,656 text segments. The median chunk length is 60 words (MAD = 20.76, range = 13–155 words).

4.2 Topic Modeling

Text Embedding : We embed text segments into a 1,024-dimensional vector space using the Voyage-AI voyage-3-large transformer-based embedding model (Voyage AI, 2025) via the VoyageAI API, accessed through Python. We select transformer-

[1] https://www.prolific.com

based models, and especially this one, due to its state-of-the-art performance on the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2022), indicating its ability to capture nuanced semantic relationships (Morris et al., 2023; Yu, 2024).

Dimensionality Reduction : As a preprocessing step for clustering, we reduce the 1,024-dimensional embedding vectors to 32 dimensions using UMAP (McInnes et al., 2020). The UMAP parameters are set to the cosine similarity metric and 15 nearest neighbors to ensure the preservation of both global and local structures in the lower-dimensional representation.

Clustering : Clustering is performed using HDBSCAN (Campello et al., 2013)) with a minimum cluster size of 3 data points, for fine-grained results. This results in the identification of 302 topics. Embeddings identified as noise ($n = 717$, 27%) are excluded from further analysis.

These steps were implemented in R (R Core Team, 2024) using the `uwot` package (Melville, 2024) for the UMAP implementation, and `dbscan` (Campello et al., 2013) for hierarchical density-based clustering.

4.3 Persistent Homology Measures

We first create a series of 27 static topic graphs (as outlined in section 3.2.2) and then use a Vietoris–Rips filtration to construct a simplicial complex from each graph (Sheehy, 2012). The filtration is based on the edge weights (cosine similarity), with simplices (edges, triangles, etc.) added as the distance threshold increases. From the resulting persistence diagrams, we extract the following topological features:

- **Betti Numbers**: β_0, β_1, β_2 represent the numbers of connected components, one-dimensional cycles (loops), and two-dimensional voids, respectively. They indicate information gaps in the graph.

- **Bottleneck Distance**: dist_B measures the maximum difference between the persistence diagrams of consecutive graphs (i.e., between graphs representing chapters n and $n + 1$). A large bottleneck distance indicates a significant singular structural change, such as the creation or filling of a large void.

- **Wasserstein Distance**: dist_W measures the average difference between the persistence diagrams of consecutive graphs. A large Wasserstein distance indicates a significant average structural change, such as a shift in the average void size.

We detrend the data using residuals from a linear model fitted to the chapter index and winsorized due to the presence of outliers. Specifically, for each feature, values below the 2.5$^{\text{th}}$ percentile are capped at the 2.5$^{\text{th}}$ percentile, and values above the 97.5$^{\text{th}}$ percentile are capped at the 97.5$^{\text{th}}$ percentile.

Network analysis is handled by the `igraph` package (Csardi and Nepusz, 2006) with `qgraph` (Epskamp et al., 2012), in R. Persistent homology and related distance measures are computed using the R packages `TDA` (Fasy et al., 2024) and `TDAstats` (Wadhwa et al., 2018). Single-threaded persistent homology calculations take approximately 100 seconds on an AMD Ryzen 7 7840U (64 GB RAM).

4.4 Generalized Additive Model

To investigate the relationship between the topological features extracted from the text and the readers' reported curiosity, we employ a Generalized Additive Model (GAM) using the R package `mgcv` (Wood, 2011). We choose GAMs for two key reasons: their ability to capture non-linear relationships and their robust options for addressing overfitting, which is crucial given our limited sample size.

We assess topological features' unique contribution to explaining variance in reader curiosity by comparing a *Null Model* (control variables: novel topics per chapter, chapter index) and a *Full Model* (control variables + topological features: Betti numbers, Wasserstein distances, Bottleneck distances).

Due to the limited number of observations ($n = 27$ chapters), our primary goal is to explore the explained deviance of the model using topological features, rather than making definitive claims about the precise functional form of the relationships. The dependent variable (DV) is the mean curiosity reported per chapter (based on $n = 49$ observations per chapter). Predictor variables (IVs) include detrended Betti numbers, Wasserstein and Bottleneck distances between chapters, and, as control variables, the number of novel topics per chapter and the chapter index number itself.

We use cubic regression splines for all smooth

functions, setting the basis dimension k to 4 for all smooth terms based on GAM diagnostics, as recommended by Wood (2017, Section 5.9). To avoid overfitting, we use Restricted Maximum Likelihood (REML) for parameter estimation with an additional penalty term ($\gamma = 1$) during model fitting. To assess the significance of the explained deviance and R^2, we employ a permutation test with 1,000 iterations, where we permute the values of the mean curiosity across chapters and refit the model.

5 Results

5.1 Distribution of the Detected Topics

HDBSCAN identifies 302 distinct clusters, i.e., topics, over the entire text. A comprehensive summary of these topics, generated using DeepSeek V3 (DeepSeek-AI et al., 2024), is provided on the online repository hosting our source code. Figure 4 shows the distribution of topics across chapters and the number of new topics by chapter, respectively. On average, each chapter contains 25 topics (SD = 5, range: 15–35). Chapter 9 exhibits the highest number of topics, while Chapter 11 contains the largest number of newly introduced topics ($n = 28$). In contrast, Chapter 14 has the fewest total topics, and Chapter 25 introduces the fewest new topics ($n = 1$).

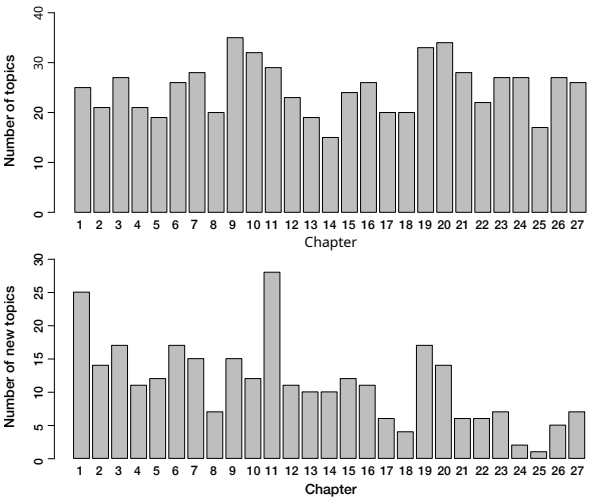

Figure 4: Total number of topics by chapter (top) and number of new topics by chapter (bottom). The x-axis represents chapter numbers, and the y-axis shows the number of topics and new topics identified within that chapter, respectively.

Visual inspection of the topic clusters across chapters reveals notable shifts in thematic content, particularly around Chapters 11 and 26, as displayed in Figure 5. These shifts correspond to key narrative transitions within the book: Chapter 11 marks the beginning of the Hunger Games, the deadly battle to be the last person standing, and Chapter 26 signifies their conclusion.

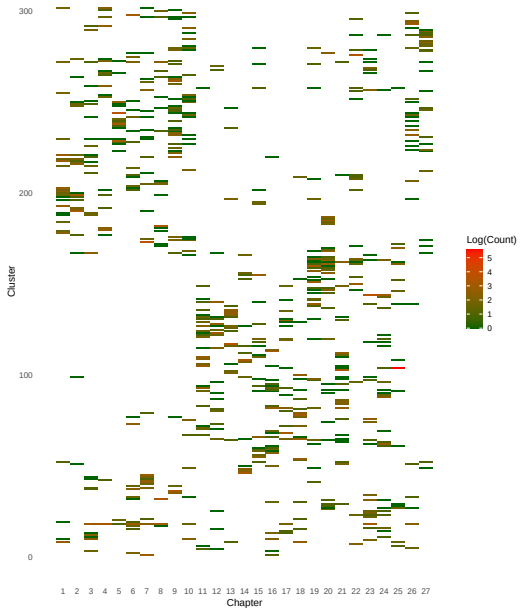

Figure 5: Distribution of topics across chapters. The chapter number is on the x-axis, and the topics are listed on the y-axis. Color indicates the frequency of each topic within a chapter, calculated as the base-2 logarithm of the number of text chunks assigned to that topic.

5.2 Description of the Extracted Topic Network

Figure 6 shows the last graph constituting our dynamic network, which contains all topics and relationships for the whole novel. It consists of 302 vertices, representing the 302 identified topics, and 778 weighted undirected edges. It exhibits an average degree of 5.15 (SD = 3.13, median = 4, MAD = 3.00, range = 1–31), a weighted diameter of 1.49 (8 when considering unweighted edges), and its degree distribution follows a log-normal law (mean-log = 1.48, SDlog = 0.56).

The small-worldness index, as defined by Humphries and Gurney (2008), is 3.40, indicating a small-world network structure. The average shortest path length (unweighted) is 3.80. A significant hierarchical tendency is observed (hierarchical clustering coefficient = 0.16, $p < .001$, (Mones et al., 2012)).

Community detection, using the Walktrap algorithm (Pons and Latapy, 2006), which showed

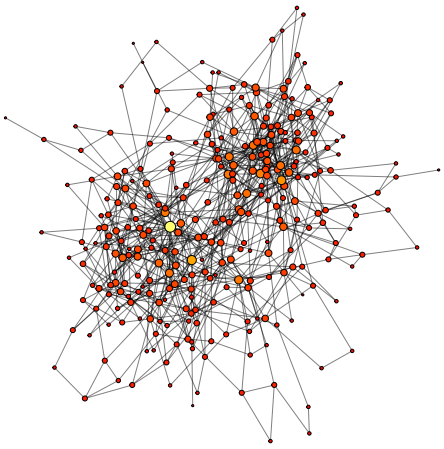

Figure 6: Topic network. The radius and the color of the vertices are proportional to their degree. The thickness of the edges is proportional to the cosine distance between the topic embeddings.

Variable	Mean	SD	Min	Max
Curiosity	69.5	3.9	60.3	77.0
Novel topics	11.2	6.3	1.0	28.0
β_0	187.0	88.9	24.0	301.0
β_1	165.7	112.3	4.0	368.0
β_2	202.6	176.0	3.0	555.0
$\text{dist}_B(\beta_0)$	0.4	0.0	0.3	0.5
$\text{dist}_B(\beta_1)$	0.5	0.2	0.3	1.0
$\text{dist}_B(\beta_2)$	0.3	0.1	0.2	1.0
$\text{dist}_W(\beta_0)$	4.5	2.5	0.5	11.6
$\text{dist}_W(\beta_1)$	6.2	2.1	2.4	9.8
$\text{dist}_W(\beta_2)$	3.2	1.9	0.5	6.4

Table 1: Descriptive statistics. The Bottleneck and Wasserstein distances are denoted by dist_B and dist_W.

good performance in previous work (Yang et al., 2016), reveals a minimum of two distinct communities within the network. A qualitative analysis of these communities reveals a strong correspondence with the narrative structure: one community primarily encompasses topics from Chapters 11–25 (the Hunger Games phase), while the other represents topics from the remaining chapters.

5.3 Persistent Homology

All derived topological features are detrended, and are shown in Figure 7. The trends observed in the three plots exhibit a resemblance to the overarching narrative structure of "The Hunger Games", where the Games themselves commence in Chapter 11 and conclude in Chapter 25. Additional descriptive statistics are shown in Table 1.

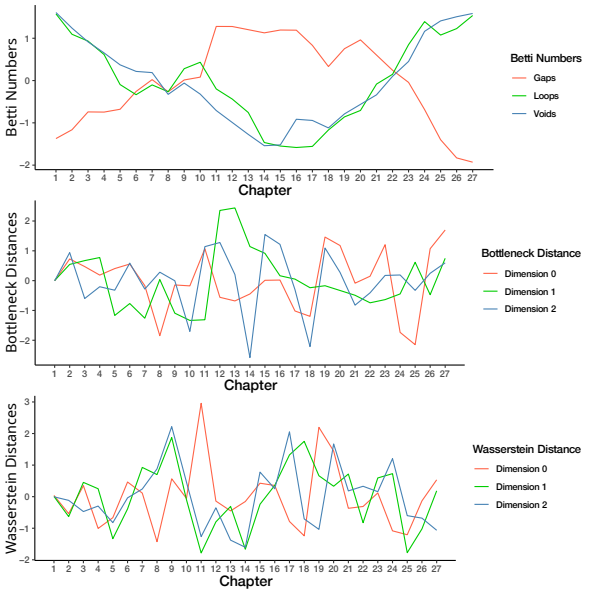

Figure 7: Betti numbers per chapter (top), Bottleneck distances between chapters (middle), and Wasserstein distances between chapters (bottom). All measures are detrended.

5.4 Generalized Additive Model

The dependent variable in the generalized additive model (GAM) is the average curiosity score per chapter (M = 69.5, SD = 3.8, range: 60.3–77.0), as shown in Figure 8. Figure 9 presents the bivariate Spearman's rank correlation matrix for all variables (Spearman, 1904).

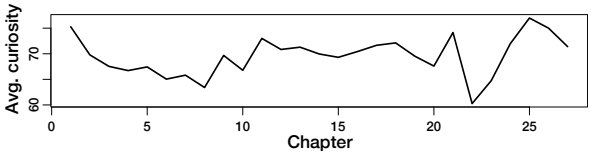

Figure 8: Average curiosity per chapter.

We compare two GAMs to assess the unique contribution of topological features towards modeling the readers' curiosity:

Null Model (Control Variables only) : It explains 23.8% of the variance and 29.7% of the deviance (permutation tests, both $p < .001$).

Full Model (Topological Features + Control Variables) : It explains 65.7% of the variance (permutation test, $p < .05$) and 72.9% of the deviance (permutation test, $p < .06$).

A likelihood ratio test comparing the full model to the null model reveals a significant improvement in model fit ($\chi^2 = 11.25$, $df = 4.7$, $p < .001$). This indicates that the inclusion of topological fea-

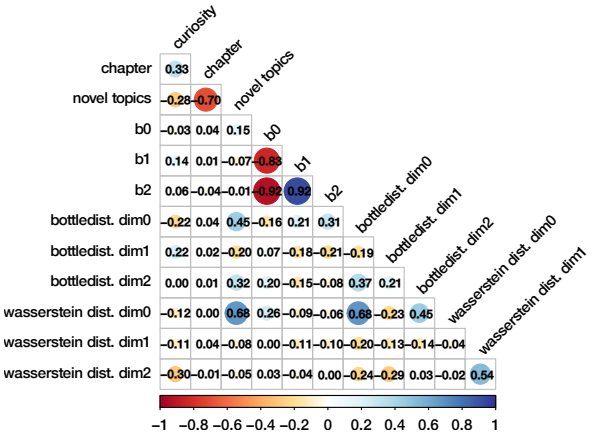

Figure 9: Spearman's bivariate rank correlations. Circle size and color are proportional to the correlation coefficient magnitude.

tures significantly improves the model's ability to explain variation in chapter-level curiosity, even after accounting for the number of novel topics and chapter index. These results show that our proposed pipeline, based on topological analysis of a dynamic topic network, can effectively model readers' curiosity ratings.

6 Conclusion

In this study, we combined existing methods into a pipeline for modeling semantic information gaps and explored their connection to reader curiosity, grounded in motivational psychology theories linking information gaps to curiosity. By constructing dynamic topic networks and using Topological Data Analysis to identify topological cavities as proxies for information gaps, we explored a distinct method from text structure analysis. Our preliminary findings demonstrate the feasibility of this approach, with statistical modeling indicating that topological features significantly enhance the prediction of reader curiosity beyond basic content and chapter progression. This proof-of-concept study provides a first step towards a quantifiable understanding of textual features associated with curiosity, potentially informing future research in computational text analysis and narrative understanding. Further research is needed to refine this pipeline, evaluate its generalizability across different text and media types, and investigate the specific contributions of various topological features to reader curiosity and engagement.

Limitations

First, our study includes a small sample size with limited measurement points focusing solely on the young adult novel "The Hunger Games". Because of this, and to avoid overfitting, we intentionally restricted our topological data analysis to simple measures like Betti numbers and distances. This may affect the robustness of our conclusions. Following this interdisciplinary proof-of-concept on one text, future work is essential to test the pipeline's robustness and applicability across diverse texts, genres (e.g., expository texts, news articles), and languages.

Second, based on our relatively homogeneous UK-based sample, results may not directly generalize to different cultural contexts, where there can be significant cultural variations in reader response (Chesnokova et al., 2017). The observed curiosity patterns are, thus, primarily reflective of this specific demographic group. Moderate reliability for mean curiosity ratings suggests that modeling individual reader differences is a key area for future work.

Third, we model texts as topic networks rather than knowledge graphs, which could limit granularity. A possibility would be to explore semantic representation with LightRAG or GraphRAG (Han et al., 2025; Guo et al., 2024).

Fourth, our network is undirected, which may overlook important learning dependencies, particularly in educational texts (Liu et al., 2012). Future work could explore directed networks to better capture the sequential progression of knowledge.

Fifth, we rely on shortest-path distances, but diffusion-based measures might better reflect how information evolves and interacts. Additionally, our focus on narrative structure constrains the applicability of our approach to raw embedding spaces. Investigating how to integrate narrative progression directly into embeddings, potentially through learning-theoretic models, should be explored. Moreover, we did not compare different embedding models, opting for a state-of-the-art option for feasibility (voyage-3-large, (Voyage AI, 2025)), so this methodological point could also be explored further. On the same note, systematic exploration of chunking parameters would be valuable in future work. Though focusing on semantic/topological features was a purposeful choice for this study, the integration of linguistic features would be a valuable direction for future research. Finally, since

our analysis is based on a narrative text, further research is needed to assess its effectiveness across different genres and domains.

References

B. Abu-Salih and S. Alotaibi. 2024. A systematic literature review of knowledge graph construction and application in education. *Heliyon*, 10(3):e25383.

M. Allaoui, M. L. Kherfi, and A. Cheriet. 2020. Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study. In *International Conference on Image and Signal Processing*, volume 12119 of *Lecture Notes in Computer Science*, pages 317–325. Springer.

M. S. Asyaky and R. Mandala. 2021. Improving the performance of hdbscan on short text clustering by using word embedding and UMAP. In *8th International Conference on Advanced Informatics: Concepts, Theory and Applications*, pages 1–6, Bandung, Indonesia. IEEE.

J. Berger, W. W. Moe, and D. A. Schweidel. 2023. What holds attention? linguistic drivers of engagement. *Journal of Marketing*, 87(5):793–809.

R. J. G. B. Campello, D. Moulavi, and J. Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, volume 7819 of *Lecture Notes in Computer Science*, pages 160–172, Berlin, DE. Springer.

K. Cao and J. Fairbanks. 2019. Unsupervised construction of knowledge graphs from text and code. In *15th International Workshop On Mining and Learning with Graphs*.

A. Chesnokova, S. Zyngier, V. Viana, J. Jandre, A. Rumbesht, and F. Ribeiro. 2017. Cross-cultural reader response to original and translated poetry: An empirical study in four languages. *Comparative Literature Studies*, 54(4):824–849.

N. H. Christianson, A. Sizemore Blevins, and D. S. Bassett. 2020. Architecture and evolution of semantic networks in mathematics texts. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2239):20190741.

S. Collins. 2011. *The Hunger Games*. Scholastic, London, UK.

G. Csardi and T. Nepusz. 2006. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

DeepSeek-AI, A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Zhang, H. Ding, H. Xin, H. Gao, H. Li, H. Qu, J. L. Cai, J. Liang, J. Guo, J. Ni, J. Li, J. Wang, J. Chen, J. Chen, Jingyang Yuan, J. Qiu, J. Li, J. Song, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Xu, L. Xia, L. Zhao, L. Wang, L. Zhang, M. Li, M. Wang, M. Zhang, M. Zhang, M. Tang, M. Li, N. Tian, P. Huang, P. Wang, P. Zhang, Q. Wang, Q. Zhu, Q. Chen, R. J. Du, Q.and Chen, R. L. Jin, R. Ge, R. Zhang, R. Pan, R. Wang, R. Xu, R. Zhang, R. Chen, S. S. Li, S. Lu, S. Zhou, S. Chen, S. Wu, S. Ye, S. Ma, S. Wang, S. Zhou, S. Yu, S. Zhou, S. Pan, T. Wang, T. Yun, T. Pei, T. Sun, W. L. Xiao, W. Zeng, W. Zhao, W. An, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, X. Q. Li, X. Jin, X. Wang, X. Bi, X. Liu, X. Wang, X. Shen, X. Chen, X. Zhang, X. Chen, X. Nie, X. Sun, X. Wang, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yu, X. Song, X. Shan, X. Zhou, X. Yang, X. Li, X. Su, X. Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Y. Zhang, Y. Xu, Y. Huang, Y. Li, Y. Zhao, Y. Sun, Y. Li, Y. Wang, Y. Yu, Y. Zheng, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Tang, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Wu, Y. Ou, Y. Zhu, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Zha, Y. Xiong, Y. Ma, Y. Yan, Y. Luo, Y. You, Y. Liu, Y. Zhou, Z. F. Wu, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Huang, Z. Zhang, Z. Xie, Z. Zhang, Z. Hao, Z. Gou, Z. Ma, Z. Yan, Z. Shao, Z. Xu, Z. Wu, Z. Zhang, Z. Li, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Gao, and Z. Pan. 2024. Deepseek-v3 technical report. *arXiv*, cs.CL:2412.19437.

N. Dvir, E. Friedman, S. Commuri, F. Yang, and J. Romano. 2023. A predictive model of digital information engagement: Forecasting user engagement with english words by incorporating cognitive biases, computational linguistics and natural language processing. *arXiv*, cs.HC:2307.14500.

S. Epskamp, A. O. J. Cramer, L. J. Waldorp, V. D. Schmittmann, and D. Borsboom. 2012. qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, 48(4):1–18.

B. T. Fasy, J. Kim, F. Lecci, C. Maria, D. L. Millman, and V.. Rouvreau. 2024. TDA: Statistical tools for topological data analysis. Accessed 2025-02-12.

M. Gerlach, T. P. Peixoto, and E. G. Altmann. 2018. A network approach to topic models. *Science Advances*, 4(7):eaaq1360.

M. Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*, cs.CL:2203.05794.

Z. Guo, L. Xia, Y. Yu, T. Ao, and C. Huang. 2024. LightRAG: Simple and fast retrieval-augmented generation. Technical report, Beijing University of Posts and Telecommunications.

H. Han, Y. Wang, H. Shomer, K. Guo, J. Ding, Y. Lei, M. Halappanavar, R. A. Rossi, S. Mukherjee, X. Tang, Q. He, Z. Hua, B. Long, T. Zhao, N. Shah,

A. Javari, Y. Xia, and J. Tang. 2025. Retrieval-augmented generation with graphs (graphRAG). *arXiv*, cs.IR:2501.00309.

M. D. Humphries and K. Gurney. 2008. Network 'small-world-ness': A quantitative method for determining canonical network equivalence. *PLoS ONE*, 3(4):e0002051.

J. Liu, J. Wang, Q. Zheng, W. Zhang, and L. Jiang. 2012. Topological analysis of knowledge maps. *Knowledge-Based Systems*, 36:260–267.

G. Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116(1):75–98.

M. M. Maslej, R. A. Mar, and V. Kuperman. 2021. The textual features of fiction that appeal to readers: Emotion and abstractness. *Psychology of Aesthetics, Creativity, and the Arts*, 15(2):272–283.

L. McInnes, J. Healy, and S. Astels. 2016. The hdbscan clustering library. Accessed 2025-02-12.

L. McInnes, J. Healy, and J. Melville. 2020. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*, stat.ML:1802.03426.

J. Melville. 2024. uwot: The uniform manifold approximation and projection (umap) method for dimensionality reduction. Accessed 2025-02-11.

E. Mones, L. Vicsek, and T. Vicsek. 2012. Hierarchy measure for complex networks. *PLoS ONE*, 7(3):e33799.

D. Moroni and M. A. Pascali. 2021. Learning topology: Bridging computational topology and machine learning. *Pattern Recognition and Image Analysis*, 31(3):443–453.

J. Morris, V. Kuleshov, V. Shmatikov, and A. Rush. 2023. Text embeddings reveal (almost) as much as text. In *Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore. Association for Computational Linguistics.

N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. 2022. MTEB: Massive text embedding benchmark. *arXiv*, cs.CL:2210.07316.

E. Munch. 2017. A user's guide to topological data analysis. *Journal of Learning Analytics*, 4(2).

S. P Patankar, D. Zhou, C. W. Lynn, J. Z. Kim, M. Ouellet, H. Ju, P. Zurn, D. M. Lydon-Staley, and D. S. Bassett. 2023. Curiosity as filling, compressing, and reconfiguring knowledge networks. *Collective Intelligence*, 2(4):1–18.

P. Pons and M. Latapy. 2006. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10(2):191–218.

R Core Team. 2024. R: A language and environment for statistical computing. Accessed 2025-02-12.

U. Schiefele. 1999. Interest and learning from text. *Scientific Studies of Reading*, 3(3):257–279.

scikit-learn developers. 2025. Comparing different clustering algorithms on toy datasets. Accessed 2025-02-12.

D. R. Sheehy. 2012. Linear-size approximations to the Vietoris–Rips filtration. In *28th Annual Symposium on Computational Geometry*, pages 239–248. ACM.

P. E. Shrount and J. L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

C. Sotirakou, D. Trilling, P. Germanakos, D. A. Sinis, and C. Mourlas. 2021. Understanding the link between audience engagement metrics and the perceived quality of online news using machine learning. *Web Intelligence*, 19(1-2):63–86.

C. Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101.

S. Tymochko, J. Chaput, T. Doster, E. Purvine, J. Warley, and T. Emerson. 2021. Con connections: Detecting fraud from abstracts using topological data analysis. In *20th IEEE International Conference on Machine Learning and Applications*, pages 403–408.

A. Uchendu and T. Le. 2024. Unveiling topological structures in text: A comprehensive survey of topological data analysis applications in NLP. *arXiv*, cs.CL:2411.10298.

Voyage AI. 2025. voyage-3-large: the new state-of-the-art general-purpose embedding model. Accessed 2025-02-11.

R. R. Wadhwa, D. F. K. Williamson, A. Dhawan, and J. G. Scott. 2018. TDAstats: R pipeline for computing persistent homology in topological data analysis. *Journal of Open Source Software*, 3(28):860.

A. Ward and D. Litman. 2008. Semantic cohesion and learning. In *International Conference on Intelligent Tutoring Systems*, volume 5091 of *Lecture Notes in Computer Science*, pages 459–469, Berlin, DE. Springer.

S. N. Wood. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(1):3–36.

S. N. Wood. 2017. *Generalized Additive Models: An Introduction with R*, 2 edition. Chapman and Hall/CRC.

Z. Yang, R. Algesheimer, and C. A. Tessone. 2016. A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, 6:30750.

J. Yu. 2024. Text embeddings to measure text topics. *Open Science Framework*.

D. Zhou, S. Patankar, D. M. Lydon-Staley, P. Zurn, M. Gerlach, and D. S. Bassett. 2024. Architectural styles of curiosity in global Wikipedia mobile app readership. *Science Advances*, 10(43):eadn3268.

X. Zhu. 2013. Persistent homology: An introduction and a new text representation for natural language processing. In *International Joint Conference on Artificial Intelligence*, pages 1953–1959.

A Inter-Rater Reliability

In order to assess Inter-Rater Reliability for mean chapter curiosity (cf. Section 4.1), we used the following formula by Shrount and Fleiss (1979):

$$ICC = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_{r,e}^2/k}, \tag{1}$$

where σ_c^2 is the variance between chapters, $\sigma_{r,e}^2$ is the interaction variance between chapters and raters, and k is the number of raters.