Tracing Thematic Change in Early English-Language Science Fiction, 1818–1930

Jonathan Gordon
Department of Computer Science
Vassar College
Poughkeepsie, NY
jgordon@vassar.edu

Abstract

How did the thematic repertoire of early English-language science fiction change as the genre consolidated between 1818 and 1930? Using a corpus of 238 public-domain texts, we apply temporally binned latent Dirichlet allocation (LDA), comparing models with and without Authorless preprocessing (which probabilistically downweights author-specific vocabulary). Cross-period topic alignments exceed a permutation null baseline, indicating continuity in topic structure over time. Full-corpus LDA can produce comparable per-topic quality, but only temporal binning enables diachronic alignment; within the binned setting, Authorless reduces author concentration and modestly increases the share of thematic topics without materially reducing coherence. Four high-continuity topic chains – centered on mobility, affect, planetary scale, and scientific knowledge – suggest a shift from earlier romantic and speculative concerns toward more consolidated technoscientific forms. These chains generate interpretable hypotheses about the literary history of early science fiction, and the workflow supports diachronic analysis in small, author-skewed corpora.

1 Introduction

Science fiction (SF) emerged as a recognizable literary mode in the nineteenth century, consolidating into a commercial genre by the early twentieth. The period from Mary Shelley's *Frankenstein* (1818) through the rise of pulp magazines in the 1920s marks a formative phase in which recurring concerns – technological change, exploration, and the nature of scientific knowledge – became legible as genre-wide patterns. Yet our understanding of how these themes evolved remains largely dependent on case studies of canonical texts rather than systematic analysis of a wider body of texts.

This paper investigates how the thematic repertoire of early English-language science fiction changed as the genre consolidated between 1818 and 1930. We address this question using latent Dirichlet allocation (LDA; Blei et al., 2003), with the goal of generating hypotheses about SF history rather than testing predetermined claims.

Topic models applied to literary corpora often recover *known structure* – such as authorship, individual works, or paratext – rather than the interpretive structure a researcher hopes to study (Rhody, 2013; Schmidt, 2013; van Zundert et al., 2022). The problem is especially acute in small corpora with high author concentration, where models may distinguish authors more readily than themes (Thompson and Mimno, 2018). To address this, we combine two strategies: we train separate LDA models on four temporal bins spanning 1818–1930 to reduce corpus heterogeneity, and we apply Thompson and Mimno's (2018) Authorless transformation to downweight author-specific vocabulary.

This paper makes two contributions. Methodologically, it shows that temporally binned models support diachronic analysis in this corpus, and that Authorless preprocessing reduces single-author dominance within that setting without materially reducing interpretability. Substantively, it offers an exploratory account of thematic change in early English-language science fiction and identifies four recurring cross-period topic chains. We treat these chains as interpretable lexical structures rather than direct equivalents of literary themes; their value lies in supporting comparative reading and generating hypotheses for literary history.

2 Related Work

In a major early study of large-scale literary topic modeling, Jockers and Mimno (2013) identified thematic clusters correlated with author gender, nationality, and publication date in a large corpus of nineteenth-century novels. Their preprocessing recommendations – notably chunking, named-entity

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 226–235

removal, and author/title filtering – inform our approach (§3.1); Sobchuk and Šeļa (2024) confirmed that such choices substantially affect resulting clusters.

Two especially relevant SF studies are Thompson and Mimno (2018) and Bologna (2020). Thompson and Mimno applied LDA to 1,206 SF novels, demonstrating that author signal poses a significant problem for literary topic modeling and proposing Authorless Topic Models as a solution. Their study addresses the SF domain but not diachronic thematic history; we adopt their mitigation strategy in service of temporal interpretation. Bologna used decade-by-decade LDA on twentieth-century SF to trace temporal lexical patterns. We extend this approach to an earlier starting point (1818–1930) and use cross-period topic alignment rather than lexicon-driven tracking.

As these and other scholars have argued, topics capture lexical regularities rather than literary themes per se – and may even recover metadata such as genre labels rather than interpretive structure (Schmidt, 2013; Rhody, 2013; van Zundert et al., 2022). We operationalize this caution by evaluating topics along three analytically separate dimensions – coherence, thematic content, and author concentration – and through close-reading vignettes that ground the four chains in primary texts. This paper brings together temporal binning, author-signal reduction, and explicit thematic evaluation for diachronic analysis of early English-language science fiction.

3 Corpus and Methods

3.1 Corpus Construction

We derive a public-domain SF corpus from a Hugging Face dataset of Project Gutenberg science fiction texts[1] and attach publication-year metadata from Momen et al. (2025), who used LLM-based inference to recover publication years for Project Gutenberg texts. In a spot check of 20 randomly sampled works, 17 were dated within ±5 years of a reference year; the remaining 3 would shift bins. (Those three were dime-novel reprints, where the metadata reflects reprint dates rather than original serialization.) Because our analysis operates on four broad bins rather than year-level trends, individual dating errors are more likely to affect bin assignments than the overall logic of periodization.

[1] huggingface.co/datasets/stevez80/Sci-Fi-Books-gutenberg

After filtering for the 1818–1930 window and removing texts for which the year estimator returned no value, our corpus contains 238 books. Texts are segmented into non-overlapping chunks of approximately 1,000 words (14,973 chunks total), following common practice for stabilizing co-occurrence statistics while keeping topics locally interpretable (Jockers and Mimno, 2013).

Table 1 summarizes the distribution across four temporal bins, chosen to balance literary-historical periods and data availability. The pre-1880 bin covers the Verne-dominated early period; later bins roughly track the rise of the scientific romance (1880–1899), Edwardian SF (1900–1914), and the early pulp era (1915–1930). Author concentration varies sharply across bins: the pre-1880 period is dominated by Jules Verne (42.5% of chunks), a share further inflated by duplicate editions with distinct volume titles that remain in the corpus. Findings from this bin therefore require particular caution; later bins have more balanced distributions.

3.2 Why Temporal Binning?

Initial whole-corpus LDA runs were unstable (mean Jaccard similarity 0.26–0.31 across random seeds, below the stability thresholds discussed by Greene et al., 2014) and remained heavily shaped by author signal despite ablations over chunk size, vocabulary filters, and deduplication. Although a higher-capacity full-corpus model later yields comparable topic quality, it cannot support the paper's central aim: cross-period alignment of thematic structure. We therefore fit separate models within more homogeneous temporal bins and trace change through alignment across adjacent periods.

3.3 Preprocessing, Modeling, and Evaluation

We apply standard preprocessing intended to reduce non-thematic structure: a document-frequency filter retaining tokens appearing in ≥10 and ≤50% of chunks (Schofield et al., 2017), supplemented by a 178-word stoplist, without lemmatization (Schofield and Mimno, 2016). Proper names are removed via a capitalization heuristic (mid-sentence capitalized words), and author/title tokens are filtered where present (Jockers, 2013). We fit LDA models using MALLET (McCallum, 2002) with 1,000 iterations, asymmetric hyperparameter optimization every 10 iterations, and a fixed random seed, with $k = 15$ topics per bin, balancing interpretability against thematic coverage at these

Bin	Books	Chunks	Authors	Top author share
Pre-1880	23	2,143	13	42.5%
1880–1899	55	3,253	42	10.0%
1900–1914	72	4,323	49	7.6%
1915–1930	88	5,254	53	15.3%
Total	238	14,973	–	–

Table 1: Corpus distribution by temporal bin.

bin sizes (2,143–5,254 chunks per bin). To assess the contribution of temporal binning itself, we also fit MALLET models to the full corpus, at $k = 15$ (matching the per-bin topic count) and at $k = 60$ (matching the total across all four bins), using the same preprocessing and settings.

To isolate the additional effect of author-signal reduction within bins, we fit two versions of each temporally binned model: standard MALLET and MALLET with Authorless preprocessing, implemented via MALLET's `DownsampleLabelWords` transform with a threshold of 0.05, which probabilistically removes tokens associated with author labels before model fitting (Thompson and Mimno, 2018). The close-reading analysis below uses the Authorless models, but we evaluate all four configurations in parallel.

Evaluation framework. We evaluate all 195 topic instances across the four configurations on three analytically separate dimensions: *coherence*, *thematic content*, and *author concentration*. We rely on manual judgment of top words together with representative chunks rather than on automated coherence metrics. Studies dating back to Chang et al. (2009) have documented systematic disagreements between automated coherence scores and human ratings of topic interpretability, and more recent work confirms that metrics such as NPMI remain unreliable proxies for human judgment in both classical and neural topic models (Doogan and Buntine, 2021; Hoyle et al., 2021, 2025). Manual inspection of top words combined with representative documents remains standard practice in DH topic modeling (Antoniak, 2022); we adopt it here.

Coherence. A topic is coded as *coherent* (yes/no) if its top words form a recognizable semantic field – that is, a set of words that co-occur because they refer to overlapping or related entities, settings, or activities. The 1915–1930 topic with top words *ship, air, water, plane, sea, deck, boat, speed, cabin, black* is coherent: the words name interrelated transportation surfaces and vehicles. By contrast, the 1880–1899 topic with top words *man, good, thing, room, time, day, made, back, put, make* is not: the list mixes high-frequency narrative and dialogue vocabulary without naming a shared domain.

Thematic content. A coherent topic is additionally coded as *thematic* (yes/no) if its semantic field picks out recurring subject matter – settings, entities, situations, or conceptual domains – rather than a linguistic mode, register, or paratext signal. Recurrence here is judged within the bin: whether the top words and high-loading chunks point to a domain that recurs across documents within the same period, independent of whether the same domain recurs across bins (the separate question of chain alignment in §3.4). The mobility topic above is thematic. A topic dominated by markers of poetic or archaic diction – for example, the 1880–1899 topic *love, thou, thy, thee, people, life, soul, heart, hath, art* – may be coherent as a linguistic register, but is not thematic in this sense; nor are topics whose top words are dominated by editorial paratext such as chapter labels or volume markers. Coherence and thematic content are intended as analytically separate dimensions: a topic can be coherent without being thematic. The dependency is one-sided – we do not code an incoherent topic as thematic, since thematic interpretation depends on a recognizable semantic field in the first place.

Author concentration. *Author concentration* is the fraction of a topic's top-5 highest-loading chunks contributed by its dominant author, ranging from 0.2 (each chunk from a different author) to 1.0 (all from the same author). It captures whether a topic's lexical pattern reflects a genre-wide regularity or one author's idiom. We treat it as a property that is reported alongside coherence and thematic content rather than collapsed into a single quality

label, since a topic can be both thematic and author-concentrated, and conflating these obscures what the Authorless transformation is actually doing.

Procedure. Coherence and thematic-content labels were assigned by a single annotator in a single pass from the top words and top-5 highest-loading chunks for each topic, conducted before chain selection so that classification was not influenced by downstream analytical choices. Coherence was judged first as a rough filter: topics judged incoherent received no thematic label, since thematic interpretation depends on a recognizable semantic field. Author concentration is computed automatically. We have not measured inter-annotator agreement; the implications of relying on a single annotator are discussed in §6.

We compare all four configurations across these three dimensions. Results appear in Tables 2 and 3.

3.4 Cross-Period Alignment

After fitting one model per bin, we align topics between adjacent periods using cosine similarity on L2-normalized topic-word distributions over their shared vocabulary (i.e., words appearing in both bins). Given a similarity matrix between topics in adjacent bins, we compute a one-to-one matching that maximizes total similarity using optimal bipartite matching (Hungarian algorithm). We use one-to-one matching as a conservative strategy, though it cannot capture topic splits or merges across periods.

We evaluate matched mean similarity against a null distribution generated from 1,000 random permutation matches per transition. Z-scores are computed as (matched mean − null mean) / null SD; p-values are derived from the standard normal distribution rather than directly from permutation counts. Results are summarized both at the level of adjacent-bin means and as multi-bin *chains* obtained by tracing matched topics across all four periods.

4 Results

4.1 Alignment Exceeds Null Baselines

Across all adjacent-bin pairs, matched topics show substantially higher similarity than expected under random matchings (Table 4). This indicates continuity in topic structure above chance, though continuity alone does not establish that the aligned topics are thematically meaningful. The weaker pre-1880 alignment (0.44 vs 0.58 for later pairs) likely reflects the bin's smaller size and high author concentration (Verne accounts for 42.5% of pre-1880 chunks).

4.2 Model Comparison: Binning, Authorless, and Full-Corpus Baselines

Table 2 compares all four configurations. Coherence rates are stable across models (80–85%), suggesting that topic quality is robust to these modeling choices. Thematic content is lowest for the full-corpus k=15 model (47%), confirming that 15 topics are insufficient for the full corpus's diversity. At k=60, however, the full-corpus model yields 62% thematic topics – comparable to binned MALLET (60%) and only modestly below binned Authorless (65%). Author concentration is actually *lower* in the full-corpus models (0.680–0.707) than in either binned pipeline (0.727–0.807), likely because no single author dominates the full corpus as strongly as Verne dominates the pre-1880 bin.

These results show that the primary contribution of temporal binning is not topic quality per se, but temporal structure: full-corpus models produce comparably thematic topics but cannot support diachronic analysis, because a single model cannot distinguish themes that emerge, evolve, or recede across periods. Binning is therefore necessary for the paper's central goal of tracing thematic change.

Within the binned setting, Authorless provides a targeted refinement. The clearest effect is on author concentration, though thematicity also rises modestly. Mean concentration falls from 0.807 to 0.727 overall (Table 5); the number of topics whose top-5 chunks all come from a single author falls from 30 to 23, and the mean number of unique authors per topic rises from 1.80 to 2.27. Authorless therefore does not mainly make topics more coherent; it makes them more representative of cross-author patterns. A topic dominated by a single author's vocabulary is stronger evidence about that author than about the genre.

Topic-quality gains are modest and uneven across periods (Table 3). Authorless improves the number of coherent-and-thematic topics in three of the four bins, with the largest gain in the pre-1880 bin (+2 topics, 53% to 67%), consistent with that bin's high Verne concentration. Concentration reductions follow a complementary pattern, strongest where author dominance is highest (Table 5).

We therefore use the binned Authorless models for the chain analysis that follows.

Model	Topics	Coherent	Thematic	Concentration
Full corpus k=15	15	12 (80%)	7 (47%)	0.680
Full corpus k=60	60	48 (80%)	37 (62%)	0.707
Binned MALLET	60	51 (85%)	36 (60%)	0.807
Binned Authorless	60	50 (83%)	39 (65%)	0.727

Table 2: Comparison of model configurations. Coherent = top words form a semantic field; thematic = coherent topic whose semantic field captures subject matter rather than register or paratext; concentration = mean fraction of top-5 chunks from dominant author (lower = more cross-author).

Bin	MALLET	Authorless	Difference
Pre-1880	8 (53%)	10 (67%)	+2
1880–1899	8 (53%)	9 (60%)	+1
1900–1914	10 (67%)	9 (60%)	−1
1915–1930	10 (67%)	11 (73%)	+1
Overall	**36 (60%)**	**39 (65%)**	**+3**

Table 3: Coherent-and-thematic topics by temporal bin.

4.3 Four Thematic Chains with Interpretable Literary-Historical Trajectories

Among multi-bin chains, we focus on four that meet two criteria. They maintain ≥0.50 mean cosine similarity across all three transitions and are interpretable as recurring content domains rather than register or author signal across all four bins. Of the 15 chains traceable across all four periods, seven meet the similarity threshold; of these, four are consistently thematic. The remaining three chains at or above the threshold are dominated by register or include paratext in one or more bins.

The close readings below illustrate how each chain's lexical signature manifests in specific highest-loading chunks rather than surveying period-wide content. The trajectories in Table 6 are interpretive labels derived from inspecting top words and top-5 highest-loading chunks together.

Chain 1: Naval/Transportation Technologies The earliest texts (Verne) treat the ship as a physical body embedded in natural forces: "the copper sheathing and the planks disappeared, reduced, no doubt, to powder" (*The Mysterious Island*, 1875). By the 1880s–1890s, emphasis shifts to velocity and strategic pursuit. In the 1900s, mobility collapses into precarious survival amid hostile spaces. By 1915–1930, the chain's highest-loading chunks treat sea, air, and space as a single operational field: "Straight down from a wisp of golden cloud a slim black speck fell toward the earth … The black flyer hung motionless" (Murray Leinster, *Murder Madness*, 1930).

Chain 2: Love, Affect, and Moral Interiority The pre-1880 anchor is dominated by Mary Shelley's *Frankenstein* (1818), where love entangles with guilt and functions as inward torment. By the 1880s–1890s, romance becomes socially mediated through family authority. The 1900–1914 period introduces philosophical stakes, while in 1915–1930 the chain's highest-loading chunks extend romantic vocabulary into cosmic registers: "Try to measure if you can, my princess, a love so vast that it draws its mate across the space between the stars" (J. U. Giesy, *Palos of the Dog Star Pack*, 1921).

Chain 3: Elemental Vastness and Planetary Forces Verne's texts frame the earth as an active, unfinished structure whose forces operate on human timescales: "It was not a question of months, but of days, it might be of hours" (*The Secret of the Island*, 1875). The 1880s–1890s shift to exploration of hidden worlds. By 1900–1914, scale becomes astronomical. The final period brings post-catastrophic landscapes among the chain's highest-loading chunks: "Slowly vegetation is creeping back upon the face of the world; but still there are vast deserts where no blade grows" (J. J. Connington, *Doomed World*, 1923).

Chain 4: Science, Knowledge, and Reality Pre-1880 frames science as moral authority and progres-

Transition	Matched	Null	Z	p
Pre-1880 → 1880–1899	0.441	0.220	7.04	$< 10^{-4}$
1880–1899 → 1900–1914	0.583	0.240	8.87	$< 10^{-4}$
1900–1914 → 1915–1930	0.586	0.252	9.06	$< 10^{-4}$
Overall	**0.537**	**0.237**	–	–

Table 4: Cross-period topic alignment. Means are cosine similarities between matched topics; null means are from random matchings; p-values are approximate (normal approximation to the permutation null).

Bin	MALLET	Authorless	Difference
Pre-1880	0.920	0.867	−0.053
1880–1899	0.867	0.693	−0.174
1900–1914	0.827	0.747	−0.080
1915–1930	0.613	0.600	−0.013
Overall	**0.807**	**0.727**	**−0.080**

Table 5: Mean author concentration by temporal bin. Lower values indicate topics that draw on a broader range of authors.

sive mastery: "So much has been done, exclaimed the soul of Frankenstein,—more, far more, will I achieve" (Mary Shelley, *Frankenstein*, 1818). By the 1880s–1890s, Camille Flammarion introduces epistemic humility: "The real nature of things entirely escapes your understanding" (*Lumen*, 1887). In 1900–1914, the top words are less lexically distinctive, but the highest-loading texts center on the epistemology of scientific observation – how evidence is marshaled and what instruments can prove. By 1915–1930, the chain's highest-loading chunks render scientific knowledge in increasingly instrumental terms: "It is not matter at all, in the ordinary sense of the word. It is almost pure crystallized energy" (E. E. Smith, *The Skylark of Space*, 1928).

4.4 Comparison with Established Genre Categories

As a qualitative point of comparison, we related selected thematic topics to established categories in the *SF Encyclopedia* (Clute et al., 2021), the standard genre-taxonomic reference. Table 7 shows mappings for topics classified as thematic; the four focal chains are marked with an asterisk.

Notable absences include time travel, aliens, and robots. These may reflect the limits of bag-of-words topics for plot-level motifs: concepts organized by narrative structure rather than lexical co-occurrence may not surface as distinct topics even when present in the corpus. Robots illustrate a different mechanism: the English translation of Čapek's *R.U.R.* (1923), which introduced the word "robot" to English-language SF, is in the corpus, but a single work cannot generate enough co-occurrence signal to form a distinct topic.

5 Discussion

5.1 Implications for SF Literary History

The four chains motivate hypotheses about how early SF reorganized recurring domains of concern across the nineteenth and early twentieth centuries. Each is consistent with – but not uniquely entailed by – the lexical evidence in Table 6 and the close readings of §4.3.

Mobility and environment. The chain is consistent with a shift from movement constrained by natural forces to movement organized by technological systems, in which the ship moves from a physically vulnerable object to a node in a larger system of signals, sensors, and weapons. The appearance of *plane* and *speed* among the 1915–1930 top words marks this shift lexically. The trajectory aligns with genre historians' emphasis on the evolution from nautical adventure (Verne) through aerial warfare to space opera.

Interiority. The chain suggests that the psychological and moral vocabulary anchored in early canonical works – notably *Frankenstein* – becomes a recurring resource: moral torment, then courtship, then a device for staging romantic devotion at cosmic scale. The lexical movement is visible across

Chain	Period	Top Words
1: Naval/ Transportation	Pre-1880	vessel, sea, ship, boat, shore, coast, wind, board, land, water
	1880–1899	ship, air, long, speed, time, miles, boat, deck, water, feet
	1900–1914	boat, ship, sea, water, deck, men, wind, shore, night, vessel
	1915–1930	ship, air, water, plane, sea, deck, boat, speed, cabin, black
2: Love/ Affect	Pre-1880	life, love, heart, father, mind, feelings, day, felt, death, thought
	1880–1899	young, girl, woman, eyes, love, father, face, mother, life, wife
	1900–1914	life, man, world, thought, love, day, heart, knew, long, night
	1915–1930	man, eyes, woman, face, love, girl, father, life, hand, voice
3: Planetary/ Elemental	Pre-1880	uncle, water, feet, volcano, lava, great, hundred, light, sea, surface
	1880–1899	water, sea, great, mountain, land, trees, lake, beneath, ice, rock
	1900–1914	water, feet, miles, sun, high, earth, great, long, hundred, mountain
	1915–1930	water, night, sun, long, day, sky, lay, wind, trees, miles
4: Science/ Knowledge	Pre-1880	life, great, men, years, time, world, man, people, country, science
	1880–1899	life, man, human, world, mind, body, nature, time, existence, power
	1900–1914	time, world, fact, made, long, work, years, life, great, planet
	1915–1930	air, space, power, light, time, machine, laboratory, metal, earth, work

Table 6: Top words for the four thematic chains across temporal bins. Trajectories: (1) Material vulnerability → Speed/pursuit → Fragile survival → Hybrid battlespace; (2) Moralized torment → Social mediation → Philosophical stakes → Cosmic devotion; (3) Natural sublimity → Hidden worlds → Astronomical scale → Post-catastrophe; (4) Rational progress → Epistemic humility → Scientific epistemology → Instrumentalized technoscience.

Our Topic	SFE Theme(s)
Naval/Transportation*	Airships, Submarines, Transportation
Love/Affect*	(not SF-specific)
Planetary/Elemental*	Mars, Moon, Astronomy, Space Flight
Science/Knowledge*	Invention, Scientists, Hard SF
War/Military	Future War, Invasion, War
Utopia	Utopias
Primitive world	Prehistoric SF, Lost Races

Table 7: Topic mapping to SF Encyclopedia themes. Asterisks indicate the four focal chains.

the top words: pre-1880 emphasizes interiority (*feelings, mind, death, thought*); 1880–1899 introduces kinship vocabulary (*girl, woman, mother, wife*); 1900–1914 is comparatively undifferentiated, mediating between the two; and 1915–1930 settles into embodied courtship terms (*eyes, face, voice, hand*). On this reading, romance is not a "non-SF" residue but a stable substrate through which SF stages novelty – implying continuities across what might otherwise look like distinct traditions in nineteenth-century scientific romance and twentieth-century pulp fiction.

Scale. The chain is consistent with a long-standing intuition in SF criticism that the genre is organized by scale shifts, and specifies one historical sequence: terrestrial geology gives way to subterranean exploration, then planetary astronomy, and finally post-catastrophic landscape. The lexical shift is visible across the four bins: *volcano, lava, surface* (pre-1880); *mountain, beneath, ice, rock* (1880–1899); *sun, earth, miles, hundred* (1900–1914); and *sky, wind, day, night, lay* (1915–1930), where distinctive geological terms have receded from the top of the distribution. The trajectory may reflect a broader tension between progress narratives and anxieties about civilizational fragility.

Science and instrumentality. The chain traces the familiar genre-historical shift from "scientific romance" to "hard SF" – from speculative philosophy and epistemic reflection toward laboratory-centered instrumentality, where physics is narrated as power over matter and perception. The lexical contrast is sharpest between 1880–1899 (*mind, body, nature, existence*) and 1915–1930 (*laboratory, machine, metal, space, power*): abstract philosophical reflection gives way to material instrumentality. This trajectory anticipates later pulp technoscience while remaining legible as an evolution rather than a sudden break.

Following Schmidt (2013), we emphasize that these patterns are shifts in vocabulary co-occurrence, not themes extracted from narrative. That they align with themes documented in traditional criticism is notable, but does not validate a naïve equation of topics with themes. The lexical evidence above shows what each hypothesis is grounded in; the same top-word trajectories nonetheless admit alternative readings, and we present them as *hypotheses suggested by topical structure* – compatible with close reading but not uniquely entailed by it. They motivate targeted genre-historical inquiry: for instance, the relationship between aviation warfare, pulp publishing venues, and the spatial settings of late-period SF.

5.2 Methodological Implications

For DH researchers working with specialized literary or historical corpora, the main lesson is practical: temporal binning is necessary when the analytical goal is diachronic, not because it produces better topics in the aggregate. Within a binned framework, preprocessing choices that suppress author signal – such as Authorless – can further shape what kinds of interpretive questions topic modeling supports. In this corpus, Authorless did not transform a poor model into a good one; it made already workable temporally binned models better suited to genre-level interpretation by broadening the author base of topics.

Genre-specific collections, author archives, and periodical runs frequently exhibit the kind of author concentration that standard topic modeling handles poorly. Authorless preprocessing should therefore be considered when the goal is thematic rather than stylistic analysis, particularly in combination with corpus segmentation: segmentation (such as temporal binning) addresses broader sources of heterogeneity, while Authorless targets the residual author concentration within each segment.

Neither intervention is without cost. Temporal binning requires post-hoc alignment to enable diachronic interpretation, and one-to-one matching is a simplification that may miss topic splitting or merging across periods. The Authorless transformation does not eliminate paratext topics or guarantee thematic coherence; it shifts distributional properties rather than substituting for interpretive judgment.

6 Limitations

Interpretive subjectivity. Topic classification and vignette selection were performed by a single annotator; different readers may label or interpret topics differently. The rubric described in §3.3 is designed to make the criteria explicit enough for replication and disagreement analysis, but inter-annotator agreement has not been measured. Two natural extensions are left to future work: a multi-annotator pass on a sample of topics with formal agreement analysis, and a parallel LLM-as-judge evaluation against the same rubric.

Concentration metric granularity. Author concentration is measured over only five chunks per

topic, so the metric takes only five possible values (0.2, 0.4, 0.6, 0.8, 1.0). Means smooth this coarseness across topics, but the per-topic measure is inherently approximate.

Temporal metadata uncertainty. Publication years include inferred values (Momen et al., 2025), which can introduce binning error. As noted above, bin-level periodization is more robust to individual date errors than year-level analysis.

Corpus representativeness and author concentration. The corpus skews late: 67% of books come from 1900–1930. Pre-1880 is dominated by a small number of authors, especially Verne (42.5% of chunks); some apparent thematic structure may partially reflect canon and availability rather than the full field of early SF. As a rough coverage check, only 9 of 25 well-known early SF works we looked for were present in the source dataset.

Model sensitivity and reproducibility. All models use a fixed random seed and identical hyperparameter settings, and we report Jaccard stability for initial whole-corpus runs (§3.2). However, we do not report formal sensitivity analysis over the number of topics k or random seeds for the binned Authorless models due to the complexity of controlling preprocessing stochasticity. Alternative binning schemes (e.g., equal-width decades) were also not tested, so results may be sensitive to bin boundaries.

Duplicate and paratext contamination. Some multi-volume editions include overlapping text; some chunks include front matter or other paratext, potentially distorting topics.

Bag-of-words constraints. Plot-level themes (e.g., time travel) and relational concepts (e.g., first contact) may not appear as distinct topics even if present in the corpus.

7 Conclusion

This paper shows that temporally binned topic modeling can support diachronic analysis in a small, author-skewed literary corpus. A comparison of four model configurations finds that full-corpus LDA at adequate capacity produces comparable per-topic quality, but only binned models enable cross-period alignment. Within that setting, Authorless preprocessing reduces author concentration and modestly increases thematicity without materially reducing coherence. Applied to early English-language science fiction (1818–1930), the workflow yields four high-continuity chains – centered on mobility, affect, planetary scale, and scientific knowledge – whose alignments exceed chance and motivate hypotheses about thematic change across the period.

Methodologically, the paper separates two preprocessing decisions that DH topic modeling often runs together. Temporal binning does not improve topic quality over a well-configured full-corpus model, but it makes diachronic interpretation possible. Author-signal reduction is a separate move: it does not enable new analyses, but it makes the resulting topics more representative of cross-author patterns. The evaluation framework introduced here – treating coherence, thematic content, and author concentration as analytically separate dimensions – supports both moves by reporting their effects separately.

The aligned chains are not direct representations of literary themes, but interpretable topic structures that support comparison, close reading, and hypothesis generation. Future work could extend the analysis to later periods, strengthen validation through multi-annotator and LLM-as-judge evaluation, and test specific hypotheses these chains motivate – for instance, the relationship between aviation warfare, pulp publishing venues, and the spatial settings of late-period SF.

Data and Code Availability

The source corpus is derived from public-domain Project Gutenberg texts and is available as a Hugging Face dataset (§3.1). Topic-level judgments (coherence, thematic content, author concentration) and analysis code will be released at `github.com/textproclab/sf-topics`.

References

Maria Antoniak. 2022. Topic modeling for the people. Blog post.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Federica Bologna. 2020. A computational approach to urban space in science fiction. *Journal of Cultural Analytics*, 5(2).

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

John Clute, David Langford, and Peter Nicholls. 2021. The encyclopedia of science fiction. Ansible Editions. Accessed 2025.

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? Re-evaluating semantic interpretability measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848, Online. Association for Computational Linguistics.

Derek Greene, Derek O'Callaghan, and Pádraig Cunningham. 2014. How many topics? stability analysis for topic models. In *Machine Learning and Knowledge Discovery in Databases*, pages 498–513, Berlin, Heidelberg. Springer.

Alexander Hoyle, Lorena Calvo-Bartolomé, Jordan Boyd-Graber, and Philip Resnik. 2025. ProxAnn: Use-oriented evaluations of topic models and document clustering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15872–15897, Vienna, Austria. Association for Computational Linguistics.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc.

Matthew L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Matthew L. Jockers and David Mimno. 2013. Significant themes in 19th-century literature. *Poetics*, 41(6):750–769.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit.

Omar Momen, Manuel Schaaf, and Alexander Mehler. 2025. Filling the temporal void: Recovering missing publication years in the Project Gutenberg corpus using LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17318–17334, Vienna, Austria. Association for Computational Linguistics.

Lisa M. Rhody. 2013. Topic modeling and figurative language. *Journal of Digital Humanities*, 2(1).

Benjamin M. Schmidt. 2013. Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, 2(1).

Alexandra Schofield, Måns Magnusson, and David Mimno. 2017. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 432–436, Valencia, Spain. Association for Computational Linguistics.

Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300.

Oleg Sobchuk and Artjoms Šeļa. 2024. Computational thematics: Comparing algorithms for clustering the genres of literary fiction. *Humanities and Social Sciences Communications*, 11:1–12.

Laure Thompson and David Mimno. 2018. Authorless topic models: Biasing models away from known structure. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3903–3914, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Joris J. van Zundert, Marijn Koolen, Julia Neugarten, Peter Boot, Willem van Hage, and Ole Mussmann. 2022. What do we talk about when we talk about topic? In *Computational Humanities Research Conference*, pages 1–18.