Predicting Topic (Co-)Occurrence Using Topic Networks Built from the Project Gutenberg Corpus

Bhuvanesh Verma, Alexander Mehler
Text Technology Lab,
Goethe University Frankfurt
{verma,mehler}@em.uni-frankfurt.de

Abstract

Although temporal topic modeling has been widely applied to scientific and legal texts, literary corpora have largely been overlooked in this regard. To address this issue, we analyze topic evolution in a subset of the Project Gutenberg (PG) corpus. We model this subset as a sequence of topic networks that capture the emergence, persistence, and interaction of thematic structures over decades. Using supervised topic representations, we predict nodes (topics) and edges (topic pairings) to forecast future topics and their co-occurrence. Our experiments demonstrate moderate to strong temporal persistence in topic connectivity patterns across three topic systems, with ROC-AUC and Average Precision (AP) values consistently above 0.85. We find that the temporal span of topic networks significantly impacts predictive performance: longer spans improve the stability and recall of topic presence, while shorter spans better capture evolving topic relationships. Overall, our findings demonstrate the predictability of topics in literary texts over time.

Keywords: Topic Evolution, Topic Network, Time-aware Networks, Temporal Autocorrelation, Project Gutenberg

1. Introduction

Literature has long served as both a mirror and a catalyst of societal change. Across centuries, written works have reflected evolving worldviews, moral values, and collective anxieties shaping how societies imagine themselves and their futures. Studying how literary topics evolve across time can therefore reveal patterns of cultural transformation, ideological shifts, and knowledge production. With the digitization of vast literary collections, such inquiries can now move beyond close reading to large-scale, data-driven analysis. Quantitative approaches in the digital humanities and computational linguistics increasingly enable researchers to trace long-term linguistic and thematic dynamics (Michel et al., 2011).

Several large-scale book corpora now facilitate such analyses, including Google Books Ngram (Michel et al., 2011), HathiTrust Digital Library (York, 2009; Jiang et al., 2021), and the Project Gutenberg Corpus (Gerlach and Font-Clos, 2020). Most previous work using these corpora has focused on statistical or descriptive analyses e.g., tracking word frequencies, stylistic trends, or sentiment trajectories (Twenge et al., 2012; Pechenick et al., 2015; Gromov and Dang, 2023) rather than modeling how thematic or topical relationships evolve structurally over time. While temporal topic modeling and dynamic co-word networks have been extensively applied to scholarly and scientific corpora to trace the emergence, convergence, and decline of research areas (Jo et al., 2011; Choudhury and Uddin, 2016; Choudhury et al., 2020), comparable frameworks for literary corpora remain underexplored, leaving open the question of how topics in literature co-evolve and interact across centuries.

With this work, we bridge this methodological gap by introducing a systematic framework for temporal topic network analysis of the temporal Project Gutenberg Corpus introduced by Momen et al. (2025). Building on its temporal dimension, we assign topics to books using both the corpus's built-in bookshelf categories (Gerlach and Font-Clos, 2020) and a state-of-the-art Dewey Decimal Classification (DDC) model (Baumartz, 2020). We then construct topic co-occurrence networks, where nodes represent topics and weighted edges represent their co-occurrence within books, aggregated over varying time period between 1700 and 1920. We investigate the autocorrelation of topic networks to evaluate their predictive continuity and define two downstream network-based learning tasks: node classification and edge classification, to assess whether past thematic structures can predict future literary relationships. To our knowledge, this work represents the first predictive, network-based temporal analysis of a large-scale literary corpus, offering new insights into the long-term evolution of thematic structures in literature.

2. Related Works

There has been extensive research on the Project Gutenberg (PG) corpus, spanning from statistical analyses of literary patterns (Gerlach and Font-Clos, 2020), semantic analysis (Egloff et al., 2019) to NLP-based subject indexing and thematic ex-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 860–869
11–16 May 2026.

ploration (Chou and Chu, 2022). While temporal topic networks have been widely studied in scholarly and scientific literature (Jo et al., 2011; Choudhury and Uddin, 2016; Choudhury et al., 2020), their application to book corpora like PG corpus remains an emerging area with substantial potential for uncovering patterns of literary and cultural evolution. More recently, the introduction of temporal dimension to the PG corpus by Momen et al. (2025), supported by advances in large language models (LLMs), has enabled the study of literary evolution over time. In the following section, we review how temporal text corpora have been modeled in the literature, focusing on methods for network construction and node and edge prediction tasks.

To model textual corpora thematically for temporal analysis, researchers often construct topic networks. A common approach is the Keyword Co-occurrence Network (KCN), where keywords are represented as nodes and connected if they appear together in the same document. For instance, Choudhury and Uddin (2016) and Choudhury et al. (2020) modeled scholarly articles using author-assigned keywords and co-occurrence frequencies as edge weights. Momeni et al. (2018), on the other hand, built dynamic semantic similarity networks of words using embedding-based cosine similarity thresholds. Both of these studies utilized clustering-based methods to identify topics within the keyword networks. Unlike these unsupervised approaches, our method employs word embedding (Mikolov et al., 2013) and Transformer (Vaswani et al., 2017) based supervised topic classification to extract topics from the corpus and construct topic occurrence networks for temporal thematic analysis.

Building on co-occurrence networks, several studies have explicitly incorporated temporal dynamics into topic networks. Choudhury et al. (2020) introduced time-aware node sets by distinguishing between current-year keywords, newly emerging keywords, and persisting keywords from previous years, and constructed multiple corresponding edge sets to capture evolving relationships. Similarly, Lin et al. (2022) modeled topic co-occurrence as a time series, applying a sliding window over co-occurrence counts to compute temporal topic similarities and build dynamic graphs, whose evolving communities reveal deeper patterns of topic interrelation over time.

Modeling time-aware topic networks naturally leads to dynamic graph analysis techniques. In graph learning, standard evaluation tasks include node-level, edge-level predictions and, graph-level prediction, such as forecasting new topics, topic co-occurrences, or the structure of the entire topic network. Jung and Segev (2022) proposed a method to identify emerging topic, represented as newly added nodes, by leveraging structural features to classify their projected future neighbors. Using 15 structural features and a binary classifier, they achieved consistent performance with accuracy and F1 scores above 0.91 across 20 datasets. Similarly, Behrouzi et al. (2020) employed structural features such as node degree, local clustering coefficient, centrality, and community score for link prediction on temporal keyword networks of scientific articles. These studies demonstrate that structural features are effective for predicting new topics and their co-occurrences in scholarly data. Following this insight, we adopt similar structural features for modeling PG corpus.

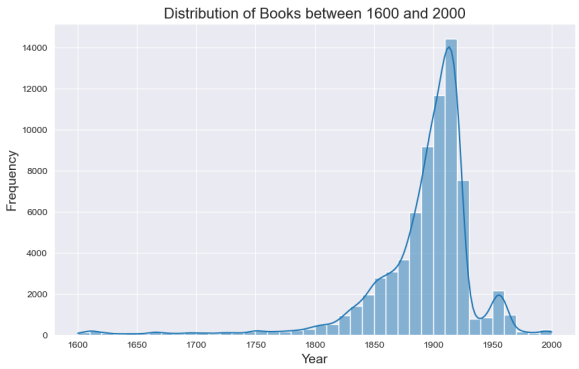

Figure 1: Distribution of books in Project Gutenberg Corpus (Momen et al., 2025) from 1600 to 2000.

3. Dataset and Methodology

We now present the dataset used in our study and the methodology employed to analyse the temporal thematic trajectory of the Project Gutenberg corpus.

3.1. Temporal Project Gutenberg (PG) Corpus

Project Gutenberg (PG) is a prominent digital library offering access to a vast collection of free eBooks. Founded in 1971 by Michael S. Hart, PG's mission is to facilitate the creation and dissemination of digital literature. As of February 2025, the library provides more than 75,000 eBooks in over 60 languages, encompassing a wide array of works such as classic novels, historical documents, and reference materials. Recently, the PG corpus was augmented with temporal metadata by Momen et al. (2025), who employed large language models (LLMs) and retrieval-augmented generation (RAG) to estimate the publication years for books.

The temporal PG corpus contains 72,978 books, of which 72,009 have recorded publication years ranging from 104 to 2023. Momen et al. (2025) applied a filtering procedure to retain only 53,774 books published between 1600 and 2000, which

are reported to have more reliable temporal estimates. Figure 1 illustrates the distribution of books over this period, which is highly skewed. To account for this, we select a subcorpus of books published between 1700 and 1920 for our topic network analysis over time. We denote this corpus by $\text{PG}_{[1700,1920]}$. This selection ensures that there are no long periods, such as the one from 1600 to 1700, with only a few books in the PG corpus.

Figure 2 illustrates the concepts of *time period*, *span*, and *window*, which are used throughout this paper. The *time period* refers to the total duration of years considered in the study. A *span* is a collection of consecutive years used to build a single topic network, while a *window* consists of multiple spans that are used together during modeling. To analyze temporal thematic patterns, we construct *topic networks*. Formally, for a given time span $[t, t+\Delta t]$, we define a graph $G_t = (V_t, E_t)$, where V_t represents the set of topics extracted from books first published within that span. An edge $(u, v) \in E_t$ is created if the topics u and v co-occur in at least one book within the corresponding span. Edge weights w_{uv} can be defined as the number of books in which the topics co-occur, i.e.,

$$w_{uv} = |\{b \in B_t : u \in \text{topics}(b) \wedge v \in \text{topics}(b)\}|,$$

where B_t denotes the set of books within the time span.

3.2. Method

We first describe the method used to study the autocorrelation of topics in books from $\text{PG}_{[1700,1920]}$. Next, we describe the predictive modeling approach based on these experiments.

3.2.1. Autocorrelation Framework

To measure thematic persistence over time, we perform two types of autocorrelation analyses on topic networks $G_t = (V_t, E_t)$.

Node-level autocorrelation: Let $x_v(t)$ denote a feature signal of node $v \in V_t$ at time t (e.g., frequency or a structural property). For a given lag τ, the Pearson-based autocorrelation is computed as

$$\rho_v(\tau) = \frac{\text{Cov}(x_v(t), x_v(t-\tau))}{\sigma^2_{x_v}},$$

where $\text{Cov}(\cdot,\cdot)$ is the covariance and $\sigma^2_{x_v}$ is the variance of the node's feature series.

Graph-level autocorrelation: Let A_t be the adjacency matrix representing the topic network G_t. We compute temporal persistence using two approaches:

1. *Pearson correlation* on global graph-level properties $g(t)$, such as average degree or clustering coefficient (Newman, 2010):

$$\rho_G(\tau) = \frac{\text{Cov}(g(t), g(t-\tau))}{\sigma^2_g}.$$

2. *Cosine similarity* between the adjacency matrices of lagged snapshots:

$$\rho_A(\tau) = \frac{\text{vec}(A_t) \cdot \text{vec}(A_{t-\tau})}{\|\text{vec}(A_t)\| \, \|\text{vec}(A_{t-\tau})\|},$$

where $\text{vec}(\cdot)$ denotes flattening the matrix into a vector; $\|\cdot\|$ is the standard Euclidean norm.

3.2.2. Predictive Models

Based on the autocorrelation analysis, we define two predictive tasks for forecasting the evolution of topic networks.

- **Node prediction**: estimating whether a topic v will appear in V_{t+1} given the current node set V_t and historical node-level signals $x_v(t)$. This task leverages structural features of the nodes as well as their temporal history.
- **Link prediction**: estimating whether an edge (u, v) will appear in E_{t+1} given the current edge set E_t, adjacency information, and historical signals. This task uses both node-level and edge-level features to capture the co-occurrence dynamics between topics.

These two tasks allow us to model and predict both the presence of individual topics and the structure of relationships between them in time windows.

4. Implementation Details

To investigate the evolution of topics within the $\text{PG}_{[1700,1920]}$, we created a dataset for autocorrelation analysis and subsequent predictive modeling. The following section outlines how the dataset was prepared and the autocorrelation and modelling steps were implemented.

4.1. Dataset Preparation

To evaluate our approach and examine the presence of temporal thematic patterns in the $\text{PG}_{[1700,1920]}$, we experimented with three different topic classification systems.

Firstly, we use the **Bookshelf Category System** (BCS), which is available by default in the $\text{PG}_{[1700,1920]}$. BCS comprises 394 predefined topic categories, which provide a baseline topical structure.

Secondly, we use the Dewey Decimal Classification (DDC) (Dewey, 1876), one of the most widely adopted hierarchical bibliographic classification systems, which consists of 10 categories at

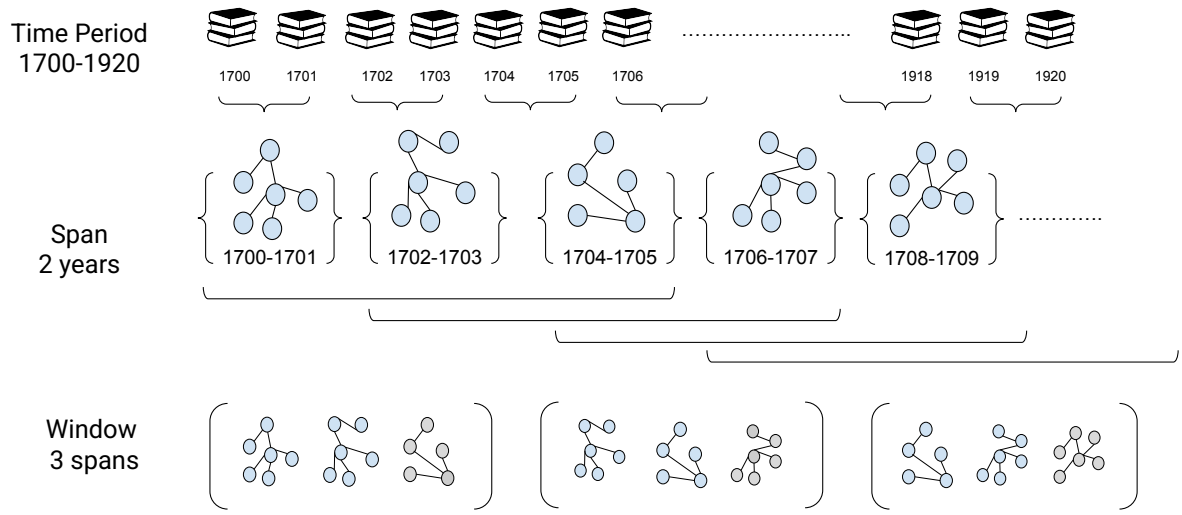

Figure 2: Illustration of time period, span, and window. The time period is the overall duration of years in the study, within which a span represents consecutive years forming one topic network, and a window comprises multiple spans used jointly for modeling. The final network in each window is used for prediction.

level 1, 100 at level 2, and 1000 at level 3. For the DDC-based categorization, we used *text2ddc* (Baumartz, 2020), a multilingual topic classifier trained to assign text to DDC categories.We mapped each sentence of each book in $PG_{[1700,1920]}$ to a second-level DDC category using text2ddc. Then, we aggregated the predictions to obtain a document- or book-level topic distribution. Since our analysis requires co-occurrence information, we focused on the three most highly predicted topics for each book. While the DDC system includes 100 second-level categories, only 78 of them were instantiated by books in the $PG_{[1700,1920]}$.

Thirdly, we incorporated the **Multilingual IPTC Media Topic Classifier** (IPTC) (Kuzman and Ljubešić, 2025), which defines 17 broad topic categories applicable across textual domains and media, including books. To reduce computational costs, we classified excerpts rather than entire books in the case of IPTC. Specifically, we randomly selected ten sentences from each book and treated each of these, along with the nine subsequent sentences, as a separate excerpt. Then, we assigned each excerpt to a single category. Although IPTC maps each excerpt to exactly one topic, aggregating across multiple excerpts enabled us to construct a category distribution for each book. This is an essential step for building the topic co-occurrence networks required for our analysis. The combined dataset with topic categories from all topic systems can be accessed here[1].

[1] https://github.com/texttechnologylab/Temporal-PG-Corpus-Analysis/tree/main/data

4.2. Auto Correlation

Our autocorrelation analysis aims to determine if thematic persistence exists across historical periods. It also examines the relevance of topics at different points in time and tracks how they emerge, fade, or disappear. To achieve these objectives, we create topic networks as temporal snapshots of the content of books from $PG_{[1700,1920]}$ We divide the corpus into consecutive spans of fixed length (e.g., 3 years), and generate a separate topic network for each span (See Figure 2).

4.2.1. Mitigating Bias in Data Sampling

One issue with this approach is that the number of books published can vary across different spans. Having a higher volume of books in one span than in another could lead to skewed topic networks and distort the autocorrelation analysis. To address this, we implemented a controlled sampling strategy. That is, to ensure that each snapshot was built from a uniform dataset, we sampled an equal number of books for each time span. This step is essential for making valid and comparable inferences across all time spans.

4.2.2. Statistical Validation

To validate our findings and establish a baseline for comparison, we incorporated a null hypothesis. This involved shuffling the time windows for the generated snapshot, which helps us determine if the observed temporal patterns are statistically significant or merely the result of random chance.

Secondly, we conducted a bootstrap experiment to calculate confidence intervals (CI), providing a

measure of the reliability of our autocorrelation estimates. To further ensure the robustness of our results, we replicated the entire experiment on the observed data. This replication allows us to account for any potential variability introduced by the random sampling process and increases confidence in the stability of our findings.

Thirdly, we calculated the autocorrelation for lags, which allows us to measure the relationship between a snapshot and its temporal predecessors. Analyzing these lags helps us to understand how topics persist over time, from short-term shifts to sustained trends.

4.2.3. Experiments

We conducted experiments at two levels. First, at the node level, we examined whether individual topics, based on their frequency or other structural properties, exhibit temporal patterns. Second, we analyzed entire topic networks using their adjacency matrices to investigate potential patterns in the co-occurrence of topics over time.

Node Level At the node level, we compute various structural properties for each topic within each time span. These properties include measures such as degree, clustering coefficient, betweenness centrality, closeness centrality, and density. We also perform a frequency-based analysis, calculating the number of books in which a given topic appears within each span. Then, we correlate these features with their corresponding lagged values to examine temporal persistence.

Graph Level At the graph level, we study global structural properties such as density, average clustering, number of communities, modularity, and average degree. These properties are correlated with lagged versions of the network to evaluate long-term stability. Beyond feature-based analysis, we perform adjacency-matrix-based autocorrelation: for each span, we compute the adjacency matrix of the corresponding topic network and measure its cosine similarity with matrices from subsequent spans. By associating each similarity score with the temporal distance (lag) between paired spans and then averaging across all pairs with the same lag, we obtain autocorrelation values that map the persistence of the network structure and topic co-occurrence over time.

In both node-level and graph-level analyses, we follow a consistent procedure. We divided $\text{PG}_{[1700,1920]}$ into spans of 10 years, each constructed from a sample of 5,000 books. We computed autocorrelation values across ten temporal lags for each topic network and the selected features. To establish statistical significance, we conducted 1,000 null experiments and 500 bootstrap experiments for each lag. The entire process is repeated 10 times to account for variability and ensure robustness.

4.3. Predictive Models

Since the autocorrelation experiments were conducted at two levels, we construct two corresponding predictive models. The first model uses information from the preceding spans to predict the presence of a node (i.e., a topic) in the subsequent time span. The second model predicts the presence of an edge, that is, the co-occurrence of topics in the subsequent time span. Combining these two models allows us to predict the structure of topic networks in the next time span.

4.3.1. Node Prediction

We use the corpus $\text{PG}_{[1700,1920]}$ also for the node prediction task. Networks are constructed for a selected span, with a maximum of 5,000 books sampled per network. We used structural properties and topic frequencies as features to model network data. We incorporate the following structural measures: node degree, strength (based on book frequencies), clustering coefficient, and betweenness centrality (Behrouzi et al., 2020). We also compute temporal features to capture the dynamics of topics. This includes the presence of a topic in earlier time spans, how long consecutive appearances last up to a given point, and how long it has been since the last occurrence. These features are related to the time-series behavior of topics, making them useful for predicting whether a topic will reappear in the near future.

We divide the sequence of networks into temporal windows to generate training data. For each temporal window of length k, we use the first $k-1$ network/s (span/s) to generate features and define the prediction task in the kth network. Note that neither the span nor the window size is fixed; thus, the model can be trained using varying historical contexts (window sizes) and network densities (spans) to predict the presence of topics.

We employ a multilayer perceptron (MLP) to model node presence. The network consists of multiple fully connected layers, each followed by ReLU activation and dropout regularization to transform node features into latent space representations. These representations are then passed through a final linear layer that applies a sigmoid activation to produce a probability distribution.

4.3.2. Link Prediction

Again, we divide the corpus $\text{PG}_{[1700,1920]}$ into fixed-length temporal spans (e.g., one year) and calcu-

Features	Bookshelves	DDC	IPTC
Density	10 (0.59@7)	4 (0.45@3)	10 (0.89@2)
Avg. Clustering	3 (0.41@3)	6 (0.36@4)	9 (0.87@1)
Num. Communities	3 (0.31@6)	0	0
Modularity	2 (0.98@7)	3 (0.23@1)	2 (0.18@5)
Avg. Degree	10 (0.90@2)	10 (0.92@1)	10 (0.896@3)

Table 1: Number of lags with significant autocorrelation ($p < 0.05$) per network feature and topic model. Values in parentheses indicate the maximum observed autocorrelation across all lags.

late node and edge features for each span. For nodes (topics), we use the features described above. For edges (topic pairings), we extract the (1) edge weight, (2) the binary existence score, (3) preferential attachment (Jeong et al., 2003), (4) temporal differences in weight, and (5) historical weight averages across past networks.

Following the node prediction strategy, we construct a dataset for training a link prediction model. We created positive and negative samples, where positive samples correspond to edges that are present in the current network. There are two sources of negative samples: (i) Hard negatives are edges that existed in previous windows but disappeared in the current one. (ii) Soft negatives are randomly sampled node pairs that never co-occurred. This sampling strategy yields a balanced set of edge instances for link prediction.

Again, we employ MLPs to perform link prediction. First, we use an MLP consisting of multiple fully connected layers that are applied to node features to produce node (topic) embeddings. Edge representations are computed by either concatenating or applying the element-wise product to the embeddings of their endpoints. These are then combined with the extracted edge features. These edge-level vectors are fed into another MLP with a non-linear ReLU activation and dropout regularization. Then, a sigmoid activation is applied to generate edge existence probabilities.

Both node and link prediction models were trained with binary cross-entropy loss using the Adam optimizer with weight decay, and a dynamic learning rate adjusted using a ReduceLROnPlateau scheduler (Al-Kababji et al., 2022). We employed early stopping based on validation loss to prevent overfitting (Yao et al., 2007). During training and evaluation, we report both ROC-AUC and Average Precision (AP). ROC-AUC measures the model's overall ability to discriminate across thresholds, while AP captures its ability to rank under class imbalance. The code and the data for all experiments is available here[2].

[2] https://github.com/texttechnologylab/Temporal-PG-Corpus-Analysis

5. Results and Discussion

We present the results of the autocorrelation and predictive modeling experiments. Since the experiments were performed at the node and graph levels, the results are split into two sections.

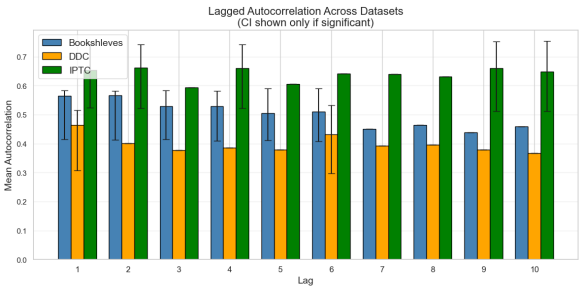

Figure 3: Autocorrelation across lags for three topic models. All bars with CI are significant.

5.1. Autocorrelation

Graph Level We observe a high degree of autocorrelation in $PG_{[1700,1920]}$ regarding the structure of the topic network and the patterns of topic co-occurrence. Figure 3 shows the average autocorrelation across all repetitions over ten temporal lags. This measure was derived from adjacency matrices by calculating the similarity between topic networks over successive time intervals. Among the three topic systems, IPTC networks exhibit the highest autocorrelation. This is likely due to the small number of topics in this model, which makes persistent connections among topics more likely over time. However, despite comprising over 350 topics, the BCS model still exhibits significant autocorrelation. This suggests that BCS topics tend to occur together throughout the examined time period. A similar trend is observed for the DDC, though the autocorrelation is comparatively lower. Despite the limited number of statistically significant lags for both the DDC and the IPTC, cyclical patterns are evident in their temporal dynamics. IPTC topics have a shorter cycle of about one decade, while DDC topics have a longer cycle of about five decades. We observe a gradual decay in topic co-occurrence for the BCS, which implies a much

longer cycle, possibly spanning centuries. These findings suggest that topic networks are remarkably persistent across decades in all three models.

Table 1 shows the autocorrelation for several structural features of topic networks. For each topic model, we report the number of lags at which a given feature's autocorrelation was statistically significant, along with the highest mean autocorrelation value across all lags. We find that the average degree exhibits strong autocorrelation at lower lags, typically between 1 and 3, across all topic models. This indicates that the connectivity of the topic networks remains relatively stable over short time intervals. Similarly, density emerges as a consistent feature over time for both BCS and IPTC, though this pattern is less pronounced for the DDC. By contrast, features related to network clustering, such as the average clustering, the number of communities, and the modularity, exhibit comparatively low autocorrelation values. This suggests that although the global structure of topic connectivity remains stable, the internal organization of topic clusters is more dynamic. Notably, community configurations undergo changes over time.

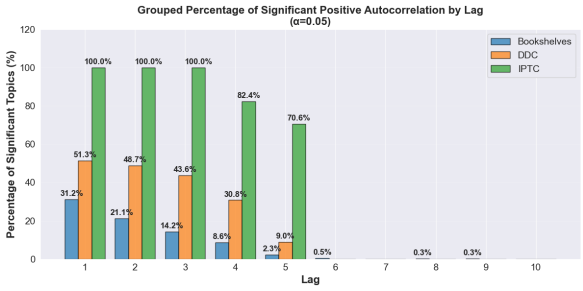

Figure 4: Fraction of topics with significant autocorrelation across three different topic systems.

Node Level We observe that, at the topic level, frequency-based autocorrelation is strong and extends across a wide range of topics for short lags (1-5). Figure 4 shows the fraction of topics exhibiting significant autocorrelation per lag. Thematic frequencies persist across a large number of topics for approximately two decades. This memory fades as the lag increases, especially after a decade.

In addition to frequency, we analyze other structural properties of topics within topic networks which mirrors the pattern observed in the frequency-based analysis. IPTC topics display the highest fraction, followed by DDC and BCS. Figure 5 shows the distribution of node degree-based autocorrelation. BCS shows moderate autocorrelation, DDC exhibits a bimodal pattern ranging from moderate to high, and IPTC shows strong autocorrelation. Consistent with the graph level, these results indicate that topics retain their structural properties over time, providing predictive insight for subsequent time spans.

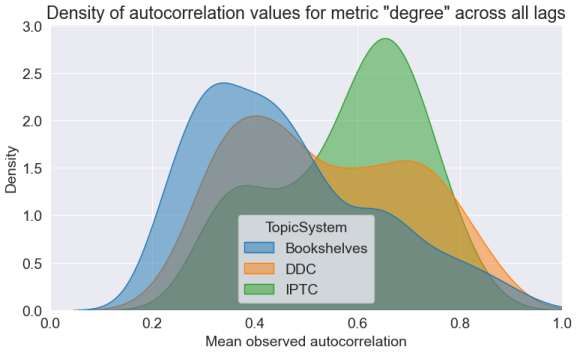

Figure 5: Distribution of mean autocorrelation values that are significant across the topic models.

5.2. Predictive Modeling

Based on the insights from autocorrelation analysis, we incorporate node-level (topic) and edge-level (topic co-occurrence) information from topic networks to predict the emergence of topics and their co-occurrence over time. Instead of setting the temporal span and window size beforehand, we perform hyperparameter tuning: We systematically vary the window size (2-10) and span size (1-10) for node and link prediction to evaluate their influence.

Dataset	Train		Val	
	ROC	AP	ROC	AP
BCS	0.894	0.828	0.886	0.936
DDC	0.962	0.974	0.955	0.992

Table 2: Train and validation ROC and AP scores.

Node Prediction Table 2 shows the results from the best parameter configurations based on average precision during training. For BCS, the optimal model was trained for 25 epochs with a single hidden layer of 256 units and a batch size of eight. The model used a 5-year temporal span and a 5-span window, meaning topic predictions for the next 5 years were made using information from the previous 20 years. This configuration achieved an average precision of 0.828 on the training set and 0.936 on the validation set. For the DDC, we observed the best overall performance across all data splits and evaluation metrics. This best-performing model was trained for 70 epochs with four hidden layers (each of size 256) and used a window of 4 spans and a span of 5 years. For IPTC, node classification was a relatively trivial task due to the small number of topics, as most nodes appeared consistently across all temporal spans. Therefore, we are

not reporting the results of node classification for this topic system.

Link Prediction In the link prediction task, we assume that the topics present in the period under consideration were known, i.e., we used gold topic data to predict edges between them.

Dataset	Train		Val	
	ROC	AP	ROC	AP
BCS	0.876	0.865	0.867	0.862
DDC	0.887	0.893	0.890	0.896
IPTC	0.944	0.987	1	1

Table 3: Train and validation ROC and AP scores.

Table 3 shows the results of the best hyperparameter configurations for all three topic systems. IPTC achieved the highest performance, reaching 100% average precision on the validation set. This is partly due to the small number of topics, which limits the possible links, and a relatively large time span of 15 years and a window size of 4. This means that the co-occurrence of topics in the next 15 years was predicted using information from the previous 45 years. Even with shorter spans of 2-3 years and windows of 2-11 spans, the model still achieved high precision (0.99 on validation, 0.95 on training).

For DDC, high precision was achieved using shorter spans and windows: the reported results were obtained by using co-occurrence information from the previous 3 years to predict the following 3 years. Similarly, for BCS, using network information from the past 2 years enabled an accurate prediction of the subsequent 2 years' topic network.

5.3. Impact of Span and Window size

It remains to analyze the effects of span and window size across all data sets and tasks.

Node Prediction We analyzed the effect of span and window on node prediction across BCS (394 categories) and DDC (78 categories), considering average precision (AP) and AUC: AUC evaluates a model's ability to rank topics that will appear in the next span higher than those that will not; AP measures the model's accuracy in predicting which topics will appear in the next span.

For BCS and DDC, span emerged as more influential than window size, which had a minimal impact. For BCS, longer spans significantly improved AP (validation correlation: 0.796). That is, access to an extended historical context is crucial for predicting the presence of a topic when there are many alternatives. However, longer spans had a negative impact on AUC (-0.979), suggesting that although frequent topics were predicted accurately, the model's ability to rank rare or emerging topics decreased. Window size had a negligible influence on both metrics.

In the case of DDC, span positively influenced AP as well, though the effect was moderate (validation correlation of 0.336). This suggests that fewer categories reduce the need for extensive historical context to predict the presence of a topic. Unlike in the case of BCS, span had a minimal negative impact on AUC (validation correlation of 0.36). This indicates that the ranking is less sensitive to long-term history when the number of topics is smaller.

Link Prediction Similar to node prediction, we analyzed the effects of span and window on link prediction. In link prediction, AP measures the model's accuracy in predicting which topic co-occurrences (edges) appear in the next time span, while AUC evaluates its ability to rank appearing edges higher than non-appearing ones.

For Bookshelves, **longer spans generally reduced both AP and AUC**, with strong negative correlations observed in training and validation sets. This suggests that aggregating many years of historical data introduces noise in networks with many possible edges, diminishing both the accuracy and ranking ability of edge predictions. **Window size had a minimal effect** overall, with only a modest negative correlation for validation AP.

For DDC, the effect of span on edge prediction was weaker, reflecting the smaller number of categories and fewer possible edges. In contrast, window size had a greater impact, especially during the validation stage, when larger windows negatively affected both AP and AUC. This suggests that, for smaller, more dynamic networks, sequentially aggregating consecutive span-based networks is critical to capturing the temporal patterns of topic co-occurrence.

Overall, we found that span plays a significant role in predicting future topics and their relationships, though its impact varies by task. This suggests that an optimal span for joint node and edge prediction must strike a balance that is effective for both predictions.

6. Conclusion

We analyzed thematic trends in book corpora to provide a basis for studies on thematic development in literary and cultural contexts. We demonstrated the predictive power of temporal topic networks at decade-long time spans using the Project Gutenberg corpus. Our experiments revealed temporal persistence at the network and node levels across three topic systems: BCS, DDC, and IPTC. These results suggest that topics and their co-occurrences can be effectively predicted based on the structural

properties of nodes and networks. We confirmed this by modeling temporal topic networks for node and link prediction. We achieved an average precision of over 0.85 and an ROC-AUC of over 0.87 across datasets. Our study also suggests that, although network density influences both node and link prediction, the sequential accumulation of historical snapshots contributes less. We provide our topic networks, based on three topic systems (BCS, DDC, and IPTC), as annotations supplementing the time-aligned version of the PG corpus from (Momen et al., 2025). In this way, we provide a resource for analyzing topic developments using historical literary corpora.

7. Ethical Consideration

This study relies exclusively on publicly available texts from Project Gutenberg, ensuring full compliance with copyright and data protection standards. Topic modeling and analysis are conducted at the corpus level, with no inferences made about individual authors or specific publications. Nonetheless, we acknowledge that the Project Gutenberg corpus may reflect biases inherent to its composition, particularly the predominance of English-language literature and Western cultural representation.

8. Limitations and Future Works

Future work can focus on discovering topic networks by jointly predicting topic nodes and their links using historical thematic data. One limitation of our study is its use of a fixed set of topics, which restricts node prediction to existing topics. However, link prediction can suggest new topic combinations that could be interpreted as new topics. Additionally, our analysis only considers the presence of topics and their relationships, not their strength of occurrence within a time span. Extending our framework to predict topic importance and relationship strength would enable us to forecast which topics or combinations thereof are likely to be most influential or popular in future time windows. Our autocorrelation results for node degree and average graph degree already indicate that this is feasible, as they show strong temporal persistence, suggesting that topic and network strength could also be predictable.

9. Bibliographical References

Ayman Al-Kababji, Faycal Bensaali, and Sarada Prasad Dakua. 2022. Scheduling techniques for liver segmentation: Reducelronplateau vs onecyclelr. In *International conference on intelligent systems and pattern recognition*, pages 204–212. Springer.

Daniel Baumartz. 2020. Automatic Topic Modeling in the Context of Digital Libraries: Mehrsprachige Korpus-basierte Erweiterung von text2ddc - eine experimentelle Studie.

Saman Behrouzi, Zahra Shafaeipour Sarmoor, Khosrow Hajsadeghi, and Kaveh Kavousi. 2020. Predicting scientific research trends based on link prediction in keyword networks. *Journal of Informetrics*, 14(4):101079.

Charlene Chou and Tony Chu. 2022. An analysis of bert (nlp) for assisted subject indexing for project gutenberg. *Cataloging & Classification Quarterly*, 60(8):807–835.

Nazim Choudhury, Fahim Faisal, and Matloob Khushi. 2020. Mining temporal evolution of knowledge graphs and genealogical features for literature-based discovery prediction. *Journal of Informetrics*, 14(3):101057.

Nazim Choudhury and Shahadat Uddin. 2016. Time-aware link prediction to explore network effects on temporal knowledge evolution. *Scientometrics*, 108(2):745–776.

Melvil Dewey. 1876. *A classification and subject index, for cataloguing and arranging the books and pamphlets of a library*. Brick row book shop, Incorporated.

Mattia Egloff, Davide Picca, and Alessandro Adamou. 2019. Extraction of character profiles from the gutenberg archive. In *Research Conference on Metadata and Semantics Research*, pages 367–372. Springer.

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126.

Vasilii A Gromov and Quynh Nhu Dang. 2023. Semantic and sentiment trajectories of literary masterpieces. *Chaos, Solitons & Fractals*, 175:113934.

Hawoong Jeong, Zoltan Néda, and Albert-László Barabási. 2003. Measuring preferential attachment in evolving networks. *Europhysics letters*, 61(4):567.

Ming Jiang, Yuerong Hu, Glen Worthey, Ryan C Dubnicek, Boris Capitanu, Deren Kudeki, and J Stephen Downie. 2021. The gutenberg-hathitrust parallel corpus: A real-world dataset for noise investigation in uncorrected ocr texts.

Yookyung Jo, John E Hopcroft, and Carl Lagoze. 2011. The web of topics: discovering the topology of topic evolution in a corpus. In *Proceedings of the 20th international conference on World wide web*, pages 257–266.

Sukhwan Jung and Aviv Segev. 2022. Analyzing the generalizability of the network-based topic emergence identification method. *Semantic Web*, 13(3):423–439.

Taja Kuzman and Nikola Ljubešić. 2025. Llm teacher-student framework for text classification with no manually annotated data: A case study in iptc news topic classification. *IEEE Access*, pages 1–1.

Weibin Lin, Xianli Wu, Zhengwei Wang, Xiaoji Wan, and Hailin Li. 2022. Topic network analysis based on co-occurrence time series clustering. *Mathematics*, 10(16):2846.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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.

Elaheh Momeni, Shanika Karunasekera, Palash Goyal, and Kristina Lerman. 2018. Modeling evolution of topics in large-scale temporal text corpora. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Mark E. J. Newman. 2010. *Networks: An Introduction*. Oxford University Press, Oxford.

Eitan Adam Pechenick, Christopher M Danforth, and Peter Sheridan Dodds. 2015. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10):e0137041.

Jean M Twenge, W Keith Campbell, and Brittany Gentile. 2012. Increases in individualistic words and phrases in american books, 1960–2008. *PloS one*, 7(7):e40181.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. 2007. On early stopping in gradient descent learning. *Constructive approximation*, 26(2):289–315.

Jeremy York. 2009. This library never forgets: Preservation, cooperation, and the making of hathitrust digital library. In *Archiving Conference*, volume 6, pages 5–10. Society of Imaging Science and Technology.

10. Language Resource References

Momen, Omar and Schaaf, Manuel and Mehler, Alexander. 2025. *Filling the Temporal Void: Recovering Missing Publication Years in the Project Gutenberg Corpus Using LLMs*.