Music Theory Timeline Explorer: Visualizing Citation and Thematic Networks

Siddharth Yalamanchili
Princeton University
sy8966@princeton.edu

May 06, 2025

1 Introduction

This project presents a plotting system for the interactive exploration of scholarly relationships in the *Music Theory Online* (MTO) corpus. Our pipeline ingests raw JSON metadata for approximately 983 papers and computes multiple graph-based relationships—including semantic similarity [Reimers and Gurevych, 2019], bibliographic coupling [Small, 1973], and author lineage—and packages the results into static JSON files.

A Dash/Plotly-based front end [Bostock et al., 2011] consumes these files to render a scrollable and zoomable timeline, in which each node represents a paper positioned by its publication date (X-axis) and a semantic-PCA coordinate (Y-axis). This layout is conceptually related to timeline citation mapping in tools like CiteSpace [Chen, 2006]. Users can toggle edge types, click to reveal local neighborhoods, and hover to inspect metadata such as title and authors—enabling progressive, uncluttered exploration of thematic and citation-based scholarly networks.

2 Approach

The architecture cleanly separates *offline* data processing (back-end pipeline) from *online* visual exploration (front-end), following best practices in modular graph-based scholarly analysis pipelines [Wang et al., 2020].

2.1 Data Preparation

The back-end is implemented as a modular Python pipeline. Each script reads or writes JSON files in the `data/` and `public/` directories:

- `load_data.py`
  - Recursively scans `data/*.json` for paper metadata.

  - Parses and normalizes fields such as `title`, `authors`, `date (YYYY-MM-DD)`, `abstract`, `keywords`, and raw `citations_raw`.
  - Builds a master list of paper dictionaries, validates dates, and slug-normalizes identifiers.
  - Outputs `nodes.json`, augmented with fields like `yPx` and `totalCitations`.

- `semantic.py`
  - Concatenates each paper's `abstract` and `keywords`.
  - Embeds text using `all-MiniLM-L6-v2` [Reimers and Gurevych, 2019], producing 384-dimensional vectors.
  - Applies PCA to extract a single scalar score z_i representing dominant semantic variation, following dimensionality-reduction methods used in bibliometric mapping [Chen, 2006].
  - Constructs a weighted undirected graph of thematic similarity using cosine-nearest neighbors ($k = 5$, threshold $= 0.6$).
  - Outputs `semantic_edges.json`, with edge weights based on cosine similarity.

- `sharedref.py`
  - Slug-normalizes all cited references.
  - Inverts the reference map to find bibliographically coupled pairs [Small, 1973].
  - Creates a weighted undirected edge for each pair of papers sharing references (weight = shared count).
  - Outputs `sharedref_edges.json`.

- `density.py`
  - Aggregates publication counts by year.
  - Outputs `year_density.json` to support temporal density visualizations in the front end.

2.2 Graph Construction

At the end of the back-end run, the pipeline produces:

$$\texttt{nodes.json} : \{\texttt{id}, \texttt{date}, \texttt{yPx}, \texttt{title}, \texttt{authors}, \ldots\},$$

$$\texttt{*edges.json}: \{\texttt{source}, \texttt{target}, \texttt{type}, \texttt{weight}\}.$$

This static serialization enables the front end to render the full graph without real-time processing.

2.3 Y-Axis Coordinate Computation and Rationale

Each node is positioned vertically to reflect its *semantic* content, enabling thematic clustering while preserving chronological order along the X-axis [Chen, 2006].

1. **Text Embedding.**
   For each paper, concatenate its abstract and keywords, then compute a 384-dimensional embedding vector $\mathbf{v}_i$ using `all-MiniLM-L6-v2` [Reimers and Gurevych, 2019].

2. **PCA Projection.**
   Apply principal component analysis (PCA) to the set $\{\mathbf{v}_i\}$ and retain the first principal component to obtain scalar values $\{z_i\}$ representing the dominant semantic variation.

3. **Min–Max Normalization.**
   Rescale each z_i to the unit interval:

   $$y_i = \frac{z_i - \min_j z_j}{\max_j z_j - \min_j z_j}$$

4. **Vertical Pixel Mapping.**
   Convert normalized values y_i into canvas pixel coordinates:

   $$yPx_i = \text{padding} + y_i \times (\text{canvasHeight} - 2\,\text{padding})$$

   This ensures effective use of the vertical display range.

5. **Overlap Avoidance (Jitter).**
   Add a small random offset to prevent overlapping nodes:

   $$\delta_i \sim \text{Uniform}(-\varepsilon, \varepsilon) \quad \text{where } \varepsilon = 3\,\text{px}.$$

   This jitter improves usability without compromising semantic integrity.

Why this Y-axis?

- Captures the *dominant semantic gradient*—papers with similar content align vertically.

- Maintains a *continuous* spatial layout, avoiding arbitrary binning.

- The first principal component often reflects features distinctive to each paper, helping ensure a unique y-coordinate per node for thematic ordering and spatial separability.

- Adds *jitter* for visual clarity without distorting semantic similarity.

3 Front-End Interaction

Users explore the precomputed graph via an interactive Dash and Plotly interface, inspired by declarative web-based visualization paradigms such as D3.js [Bostock et al., 2011]. The layout consists of a control panel on the left and a scrollable graph panel on the right.

- **Edge-Type Selector:** A radio-button widget lets the user choose exactly one edge type (semantic, shared-ref, or lineage). That choice determines which edges appear when a node is clicked or when "Show All" is pressed.
- **Node Clicks:** Clicking on a paper node displays only the edges of the selected type that connect to that node, hiding all other edges and nodes to focus on the local neighborhood.
- **Show All:** The "Show All" button renders every edge of the chosen type while keeping all nodes visible, providing a global overview of that relationship.
- **Reset View:** The "Reset View" button clears all edges and returns to the initial state, where all nodes are visible and no edges are drawn.
- **Search Highlight:** A text search field highlights nodes whose title or author names match the query; matching nodes are recolored but never hidden.
- **Hover Tooltips:** Hovering over any visible node brings up a floating metadata box showing the paper's title and authors, regardless of current filters or edge visibility.
- **Canvas Layering:** Nodes are always drawn on top of edges so that click and hover interactions target the node markers. Edges have hover disabled to avoid interference with node interactions.

This design supports both focused exploration of individual papers' relationships and broad surveys of thematic and citation patterns over time.

4 Conclusion and Future Work

This project offers a new perspective on how music theory as a discipline has developed, diversified, and organized itself over time. By visualizing not only who cited whom, but also how authors think alike, cite similarly, and trace intellectual lineages, we present a richer account of scholarly influence.

4.1 A Story Told in Edges

Three edge types in particular reveal distinct dimensions of academic development [Small, 1973, Chen, 2006]:

- **Semantic edges** uncover latent thematic relationships—papers that may never cite each other but address similar problems or share theoretical frameworks. This enables the discovery of parallel lines of inquiry and thematic convergence within the field.

- **Shared reference edges** highlight how certain foundational works unite otherwise disparate strands of research. When papers draw on the same sources, it reflects a shared intellectual foundation—even in the absence of direct citation.

- **Author lineage edges** trace mentorship, teaching, and intellectual inheritance from one scholar to another, showcasing how ideas are passed on and reshaped across generations.

Together, these perspectives shift the focus from isolated citation metrics to a more nuanced view of how knowledge is built—collaboratively, thematically, and historically. For music theory, a field that blends analytical rigor with cultural and philosophical inquiry, this layered representation helps reveal the hidden structure of its scholarly evolution.

4.2 Leveraging Advanced LLMs for Enhanced Similarity

Our current semantic axis is based on a lightweight SentenceTransformer model (`all-MiniLM-L6-v2`) with 1D PCA projection [Reimers and Gurevych, 2019]. Future work can extend this with deeper, context-aware techniques:

- **Use powerful LLM embeddings:** Replace or augment MiniLM with embeddings from larger models such as GPT-4 [OpenAI, 2023], fine-tuned on scholarly text. These models capture deeper syntactic and semantic context.

- **Contextual citation embeddings:** Embed not just abstracts, but also citation sentences and discussion passages to reflect how papers engage with each other in scholarly discourse.

- **Dynamic, on-demand re-embedding:** Enable API-driven front-end re-embedding for user-selected subsets, allowing customization of similarity thresholds and exploration of newly added papers.

- **Hybrid graph metrics:** Integrate semantic embeddings with graph-theoretic measures (e.g., shortest-path citation distances) to reflect both topical and structural proximity [Wang et al., 2020].

By incorporating advanced LLMs and richer embedding strategies, the system can reveal subtler patterns—emerging subfields, parallel theoretical trajectories, and tightly knit citation communities [Wang et al., 2020]—that go beyond what PCA can offer.

Source Code and Data Repository:
`https://github.com/namaste-world/MTO_Plot`

References

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12): 2301–2309, 2011.

Chaomei Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3):359–377, 2006.

OpenAI. Gpt-4 technical report, 2023. URL `https://cdn.openai.com/papers/gpt-4.pdf`.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. ACL, 2019.

Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, 1973.

Kai Wang, Chao Shen, Jun Liu, and Yong Gao. Graph embedding for scholarly networks: Methods, applications, and challenges. *IEEE Access*, 8:206383–206399, 2020.