Multi-Scale Topic Manifold Learning for Understanding Interdisciplinary Collaborations in Co-Author Networks

Anonymous Authors[1]

Abstract

A multi-scale model of topic relationships and trends is introduced for co-authored document corpora. Hellinger-PHATE topic space embeddings are integrated with hierarachical random graph models for unique visual analyses of author collaborations. Co-author network links are ranked using maximum likelihood estimates, under assumptions of a mixed-membership stochastic block model. In addition, topic manifold embeddings are combined with ranked interdisciplinarity scores for both authors and documents, bridging multiple topic clusters. Empirical visualizations demonstrate the power of interpretable geometric and probabilistic models for unsupervised multi-modal analysis of evolving text corpora and co-author networks.

1. Introduction

Modern unsupervised learning offers powerful methods for uncovering hidden patterns and structures in unlabeled data. However, most algorithms are constrained to uni-modal data, which limits their effectiveness in environments with nonlinear relational data, such as networks. These environments often require extracting multi-scale patterns (local, meso, and global) from noisy, multi-modal data, motivating a new approach.

Topic modeling automates the discovery of semantic patterns in document corpora. Early methods like TF-IDF (Sparck Jones, 1972), LSA (Deerwester et al., 1990), and NMF (Lee & Seung, 1999) reduced dimensionality using linear algebra, while probabilistic models such as pLSA (Hofmann, 1999) and LDA (Blei et al., 2003) introduced document-topic mixtures. Embedding advancements like Word2Vec (Mikolov et al., 2013) and transformer models such as BERT (Kenton & Toutanova, 2019) and SciBERT (Beltagy et al., 2019) have further refined text representations. Domain-specific adaptations, such as NukeLM (Burke et al., 2023), leverage transfer learning for specialized tasks.

Despite these advancements, integrating text and network data remains a significant challenge. LDA's interpretable probabilistic framework and tools like LDAvis (Sievert & Shirley, 2014) enable effective topic discovery and visualization, but they lack the capacity to incorporate relational structures inherent in networks. Recent geometry-driven ensemble methods (Wang et al., 2021), leveraging Hellinger distance (Hellinger, 1909) and PHATE (Moon et al., 2019), extend LDA's temporal and geometric utility, yet they still focus on uni-modal data. Meanwhile, graph learning techniques excel at modeling relational structures but rarely incorporate text semantics, creating a gap in the joint analysis of text and network data.

Efforts to bridge this gap often rely on task-specific models or metadata that is not universally available. For example, LDA-based recommender systems (Hwang et al., 2017) and diffusion network analytics (Zhang et al., 2023) reveal collaboration patterns but do not scale seamlessly across disciplines or domains. This limitation underscores the need for methods that are not only interpretable and computationally efficient but also capable of analyzing multi-modal and multi-scale patterns in a unified framework.

To address these challenges, we introduce multi-scale topic manifold learning (MSTML), which combines geometry-driven ensemble topic modeling (Wang et al., 2021) with hierarchical random graphs (Clauset et al., 2008). MSTML bridges the gap between text and network analysis by providing an unsupervised, interpretable probabilistic model that captures multi-modal and multi-scale patterns. Unlike existing approaches, MSTML requires only text documents and author lists, eliminating dependencies on metadata. Applications to arXiv documents reveal insights into interdisciplinary collaborations and author trends. Section 2 introduces the theoretical foundations, Section 3 outlines MSTML, and Section 4 details case studies, followed by conclusions in Section 5.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. BACKGROUND

2.1. Topic Modeling

2.1.1. Latent Dirichlet Allocation

Topic modeling aims to uncover thematic structures in document corpora. LDA is a generative probabilistic topic model and one of the most commonly-cited techniques (Blei et al., 2003). Typically, LDA is diagrammed as in Figure 1.

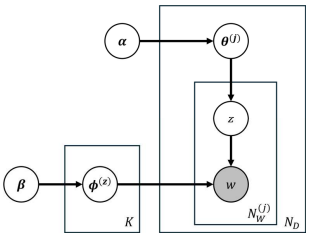


Figure 1. LDA Bayesian network model. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are Dirichlet priors for multinomial vectors $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. There are N_D total document-topic distribution vectors, $\{\boldsymbol{\theta^{(j)}}\}_{j=1}^{N_D}$. $\boldsymbol{\theta^{(j)}}$ encodes the distribution of document j over K word-frequency vectors (topics), $\{\boldsymbol{\phi^{(k)}}\}_{k=1}^{K}$. Topic vectors are multinomial distributions over vocabulary $\mathcal{V}$. Word, w, are directly observable (gray shaded). There are $N_W^{(j)}$ words in document j.

In LDA, each word in a document is an outcome of a generative process which randomly selects a topic z, then a word w. Topic z is first sampled according to multinomial distribution $\boldsymbol{\theta^{(j)}}$. Second, w is sampled from the vocabulary, according to the multinomial distribution $\boldsymbol{\phi^{(z)}}$. This repeats for all $N_W^{(j)}$ words in document j, and for all documents $i \in \{1, \ldots, N_D\}$. LDA uses Bayesian inference to learn the vectors of interest, $\{\boldsymbol{\theta^{(j)}}\}_{j=1}^{N_D}$ and $\{\boldsymbol{\phi^{(k)}}\}_{k=1}^{K}$.

2.1.2. Text Preprocessing

Large vocabulary sizes are computationally challenging (Manning et al., 2008). Therefore, connective stop words that do not contribute semantic information, including conjunctions and articles, should be removed prior to model training. Variations of the same word can also be joined into a common root form, which is called stemming. Lemmatization further combines semantically-similar terms, like "better" and "good." Such operations may extend to abbreviations and acronyms.

Techniques for reducing the vocabulary often utilize term frequency thresholds. Filtering must balance the trade-off between noise and topic granularity (Lu et al., 2017; Maier et al., 2021). Terms that appear only once in the corpus can be safely removed, but terms that appear more than once, and high-frequency terms, add to document discriminative capacity. The same can be said for increasing the number of topics in LDA (Lu et al., 2017).

Sievert & Shirley (2014) developed a nice linear embedding and visualization tool, called LDAvis, which used the concept of term *relevancy*, defined in Equation (1). $P(w \mid k)$ is the probablity of term w given the topic k, and $P(w)$ is the marginal probability of the term w across the entire corpus. $\lambda \in (0, 1)$ is a weight hyperparameter. Term relevancy balances frequency within a topic against exclusivity to that topic, providing a basis for vocabulary filtering by ranking terms using relevancy scores from an auxiliary model.

$$r(w, k \mid \lambda) = \lambda \log P(w \mid k) + (1 - \lambda) \log \left(\frac{P(w \mid k)}{P(w)} \right) \tag{1}$$

2.1.3. Information-Geometric Topic Models

Information geometry applies differential geometric concepts to probability distributions. Unlike standard Euclidean vectors, the L_2 distance is ineffective for comparing multinomial vectors, such as $\boldsymbol{\theta}$ or $\boldsymbol{\phi}$ from LDA (Nielsen, 2020; Sun & Marchand-Maillet, 2014; Wang et al., 2021).

The Fisher information defines a Riemannian metric on statistical manifolds, including the manifold of LDA topic vectors. Due to its computational simplicity, the Hellinger metric is commonly used as an approximation of the Fisher information. Given two multinomial vectors, $\boldsymbol{p}$ and $\boldsymbol{q}$, the Hellinger metric is defined according to Equation (2):

$$H(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{n} (\sqrt{\boldsymbol{p}_i} - \sqrt{\boldsymbol{q}_i})^2} \tag{2}$$

The Hellinger distance exhibits useful properties of being bounded between 0 and 1 and symmetric between $\boldsymbol{p}$ and $\boldsymbol{q}$.

2.2. Manifold Learning

Manifold learning is a sub-set of methods for dimension reduction of high-dimensional data. Unlike linear methods like principal component analysis (PCA) (Pearson, 1901), manifolds are assumed to be globally non-linear but locally linear. Well known algorithms include locally linear embedding (LLE), t-SNE and multi-dimensional scaling (MDS) (Roweis & Saul, 2000; van der Maaten & Hinton, 2008; Kruskal, 1964).

2.2.1. Fast Approximate Nearest Neighbors

Semantic vector search is a key application of topic modeling, leveraging fast retrieval of similar high-dimensional vectors. Retrieving similar vectors is also a foundational step in manifold learning due to the construction of nearest neighbor graphs to capture data geometry. FAISS, a popular open-source library, facilitates these processes by leveraging product quantization (PQ) to reduce storage requirements and perform fast approximate nearest neighbor

(ANN) search, enabling scalability to millions of vectors (Jegou et al., 2010; Johnson et al., 2019).

Given a vector $\boldsymbol{x} \in \mathbb{R}^d$, Product quantization splits $\boldsymbol{x}$ into m sub-vectors, $\boldsymbol{x} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_m}]$, where $\boldsymbol{x_i} \in \mathbb{R}^{d/m}$. Each sub-vector $\boldsymbol{x_i}$ is approximated using centroids from a codebook, $\boldsymbol{x_i} \approx \boldsymbol{c_{i,j}}$, reducing dimensionality while preserving approximate geometry. FAISS asymmetrically quantizes only database vectors, leaving query vectors uncompressed for search accuracy. Efficient search in the probability simplex is particularly difficult (Krstovski et al., 2013). However, the Hellinger distance can be readily used with FAISS by taking the element-wise square root of each vector. Pairwise distance computations scale as $O(n^2)$, but FAISS reduces this complexity to $O(n \cdot d)$.

2.3. Authors and Co-Authorship

2.3.1. Co-Author Networks

A co-author network is defined by observable author lists associated with each document in a corpus. When multiple authors share approximately the same name, author disambiguation is critical for subsequent analyses of co-authorship patterns. A common method is to use a fuzzy matching algorithm or additional meta-data (Hwang et al., 2017). Once author disambiguation has been performed, co-author networks are defined in a straightforward manner:

Definition 2.1. A **co-author network** is a graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, where: (a) $\mathscr{V}$ is the set of vertices (nodes), where each vertex $v \in \mathscr{V}$ represents an author, and (b) $\mathscr{E} \subseteq \mathscr{V} \times \mathscr{V}$ is the set of edges (links), where each edge $e = (u, v) \in \mathscr{E}$ indicates that authors u and v have co-authored at least one document together. The edge set $\mathscr{E}$ is typically undirected, so $(u, v) \in \mathscr{E} \implies (v, u) \in \mathscr{E}$. The degree $\deg(v)$ of a vertex $v \in \mathscr{V}$ is the number of co-authors of author v; equivalently, the number of incident edges.

2.3.2. Interdisciplinarity

Interdisciplinarity refers to the cross-collaboration of scientists or authors from multiple fields. There is an increasing interest in understanding the impact of interdisciplinarity in scientific works (Okamura, 2019; Ullah et al., 2022). Depending on the metric, interdisciplinarity has been shown to be positively correlated with increased research impact, based on citation metrics. Typical measurements of interdisciplinarity utilize manually-labeled document categories, but MSTML seeks to avoid using additional meta-data.

2.3.3. Link Prediction and Evaluation

Link prediction is a well-researched problem in network science. Given a co-author network $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, the task is to predict future or missing links (u, v) that are not observed in $\mathscr{E}$. Standard methods are derived from the common neighbors assumption, that the likelihood of $(x, y) \in \mathscr{E}$ increases for each additional common neighbor between x and y (Adamic & Adar, 2003; Liben-Nowell & Kleinberg, 2003; Nassar et al., 2019). More recent work typically uses graph embeddings or graph neural networks (GNNs) (Zhu et al., 2023; Kumar et al., 2020).

Link evaluation is a related problem, with the goal of assessing the likelihood of observed and unobserved links. Topological properties and local assortativity within a network can be used to model and infer likelihoods. MSTML prioritizes link evaluation over prediction, employing a hierarchical random graph model (Clauset et al., 2008) to assign likelihoods.

2.4. Multi-Scale Learning

Multi-scale learning captures local, meso, and global patterns. In MSTML, hierarchical topic relationships are formed by agglomerative clustering, which iteratively merges the most similar entities, or vectors. This produces a dendrogram tree of intuitive visual cluster relationships. Unlike other binary trees, dendrogram nodes, m, are each associated with a dissimilarity d, where the left and right child nodes m_l, m_r are d-dissimilar. Merge heights in the dendrogram represent this visually. Analyzing data at a particular scale begins by truncating the dendrogram.

Dissimilarity depends not only on the metric but also the clustering linkage type. Common linkage methods are: single, complete, average, and Ward. MSTML uses Ward's method by default, minimizing total within-cluster variance by merging clusters with the smallest increase in squared differences. For clusters C_i, C_j, with centroids $\boldsymbol{\mu_i}, \boldsymbol{\mu_j}$, Ward's linkage uses dissimilarity function,

$$d(C_i, C_j) = \frac{|C_i| \cdot |C_j|}{|C_i| + |C_j|} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2. \tag{3}$$

2.4.1. Hierarchical Random Graphs

Clauset et al. (2008) developed a probabilistic hierarchical link prediction algorithm which is capable of both link prediction and evaluation. This model, known as hierarchical random graphs (HRG), forms a dendrogram to model assortative or dis-assortative structures in networks. The dendrogram structure is sampled according to a Markov-chain Monte Carlo (MCMC) process, based on the observed network. In the HRG model, we associate a probability p_m with each internal node m. Given a pair of vertices $u \in \mathscr{V}, v \in \mathscr{V}$, the probability of a connecting edge is $P((u, v) \in \mathscr{E}) = p_{u,v} = p_m$, where m is the lowest common ancestor of u, v in $\mathscr{D}$. The combination of the dendrogram $\mathscr{D}$ and internal node probabilities $\{p_m\}$ constitutes a hierarchical random graph.

Definition 2.2. Let $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ be a graph with $n = |\mathscr{V}|$ vertices. A **dendrogram** $\mathscr{D}$ is a binary tree with n leaves $\{m_1, m_2, \ldots, m_n\}$, each leaf corresponding to a vertex in $\mathscr{G}$. $\mathscr{D}$ also contains $n-1$ internal nodes $\{m_{n+1}, m_{n+2}, \ldots, m_{2n-1}\}$, each of which corresponds to a group of vertices which are descendants of that internal node.

The HRG model (Clauset et al., 2008) is a variation on the classical $\mathscr{G}(n,p)$ random graph model of Erdős–Rényi. The presence or absence of any edge is considered independent of the rest. The HRG model modifies the $\mathscr{G}(n,p)$ model by assuming edge probabilities are inhomogeneous, where the inhomogeneities are controlled by the topology of $\mathscr{D}$. This mirrors the development of stochastic blockmodels in network science, which also treat links as independently generated with inhomogeneous probabilities (Holland et al., 1983; Airoldi et al., 2008; Karrer & Newman, 2011; Lu & Szymanski, 2019).

For HRG models, the likelihood that a specific model explains the observed network is based on Equation (4), which treats the HRG model as a collection of binomial random variables. Each internal node is associated with a binomial random variable with success probability p_m:

$$\mathscr{L}(\mathscr{D}, \{p_m\}) = \prod_{m \in \mathscr{D}} p_m^{\mathscr{E}_m} (1-p_m)^{L_m R_m - \mathscr{E}_m} \qquad (4)$$

In Equation (4), $\mathscr{E}_m$ is the number of edges in $\mathscr{G}$ whose endpoints have internal dendrogram node m as their lowest common ancestor. L_m and R_m are the numbers of leaves in the left and right subtrees of the dendrogram node m, respectively. When the dendrogram is fixed, the probabilities that maximize the likelihood $\mathscr{L}(\mathscr{D}, \{p_m\})$ are given by

$$\left\{ \hat{p}_m = \frac{\mathscr{E}_m}{L_m R_m}, \forall m \right\}. \qquad (5)$$

Dendrogram topologies are sampled and then evaluated for likelihood, according to Equation (4). The internal node probabilities, $\{p_m\}$, encode the likelihood to observed or unobserved links in the network. Under the HRG model, observed links with low likelihoods can be classified as anomalous.

3. METHODS

3.1. MSTML Model and Algorithm

MSTML integrates topic manifold learning (Wang et al., 2021) with hierarchical network models (Clauset et al., 2008). The core method fits a dendrogram, parameterized by internal node probabilities, $\{\mathscr{D}; \{p_m\}\}$, to a co-author network $\mathscr{G}$. The authors (nodes) in $\mathscr{G}$ are represented by embeddings derived from an LDA topic model ensemble.

MSTML maps LDA-derived topic vectors $\{\boldsymbol{\phi}^{(k)}\}$ to dendrogram leaves. These topic vectors reside on the probability simplex, $\Delta^{\nu-1} \subset \mathbb{R}^{\nu}$, where ν is the vocabulary size. The MCMC process of the HRG model, which learns the dendrogram topology, is replaced by agglomeratve clustering, using the Hellinger distance (Equation 2). FAISS provides fast approximate k-nearest neighbors for fine-tuning the learned manifold. Figure 2 illustrates an example topic dendrogram model.

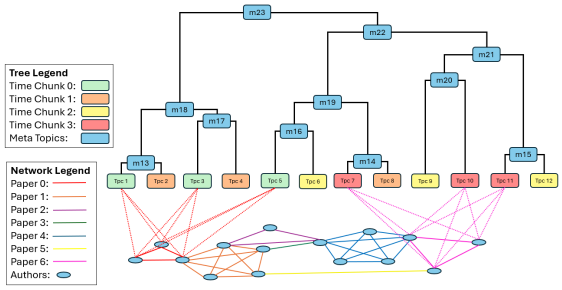


Figure 2. The topic space dendrogram links chunk topics (multi-colored rectangles), meta-topics (m13-m23), and authors (blue circles). Authors are embedded as multinomial vectors derived from the LDA ensemble. Several example links between the dendrogram and co-author network indicate this relationship.

3.2. Topic Model Ensemble

Instead of one LDA model, MSTML employs an ensemble learning approach. The corpus is split into uniform time chunks, and LDA is trained on each sub-corpus $C_1, C_2, \ldots, C_T$ independently. Each LDA model uses uniform Dirichlet priors, $\boldsymbol{\alpha} = \mathbf{1}, \boldsymbol{\beta} = \mathbf{1}$. A key challenge with LDA lies in choosing the number of latent topics. Prior literature is mixed on the appropriate number of topics, K. As described in Gan & Qi (2021), selecting K is usually based on: avoiding topic duplication, high isolation between topics, and repeatability. These factors are measured by evaluating topic perplexity, isolation, stability, and coincidence. We have chosen to scale K as an affine function of the number of documents per time chunk. The goal is to encourage topic trend continuity while detecting new, emerging topics.

The document set is ordered by publication date, prior to splitting. This results in a sub-corpus for each time chunk, C_t, where $t \in \{1, \ldots, T\}$ is the chunk index. We then apply a corpus chunk smoothing process by exponential subsampling, controlled by a decay parameter γ, leading to subsampled sub-corpora $\{C'_1, C'_2, \ldots, C'_T\}$. In this notation, C'_1 is the corpus created by sampling $\gamma^{(\tau-t)-1}$ fraction of the documents from the $C_{\tau-t}$ chunk. We use the notation C_t instead of C'_t for simplicity, despite the fact that we are often referring to the smoothed corpus chunks.

Topics are derived from the temporally-chunked sub-corpora, then clustered into meta-topics. Temporal smooth-

ing improves continuity in the topic geometry as explained in Wang et al. (2021). The resultant, smoothed manifold allows for interpretable traversal and alignment with co-author network trends. The ensemble approach prevents overfitting on any given LDA model. Simultaneously, the odds improve for capturing niche topics that would be diluted by a single, global topic model.

3.3. Author Embeddings

3.3.1. Author-Topic Barycenters

Trained models produce topic vectors $\{\boldsymbol{\phi^{(k)}}\}$ and document-topic distribution vectors $\{\boldsymbol{\theta^{(j)}}\}_{j=1}^{N_D}$ Each document-topic distribution vector spans only $K^{(t)}$ topics of its source time chunk, so $\boldsymbol{\theta^{(j)}} \in \Delta^{K^{(t)}-1}$. We seek a map for each author $u \in \mathscr{V}$ to a representation $\boldsymbol{\psi^{(u)}}$, over all corpus topics. Specifically, we seek a multinomial vector because of the advantages of working within the same representation space as in LDA.

Each author, $u \in \mathscr{V}$, is associated with a corpus set $C^{(u)}$, the set of documents authored by u. Let $C_t^{(u)} \subset C^{(u)}$ be the subset published during time chunk t. For each document j in $C^{(u)}$, let the associated author list be $\mathscr{V}^{(j)} \subset \mathscr{V}$. Recall that each document is associated with distribution $\boldsymbol{\theta^{(j)}}$. For author u, embedding vector $\boldsymbol{\xi^{(u)}}$ is a weighted mean of the document-topic vectors in $C^{(u)}$, where the weights are determined by number of co-authors (Equation (6)).

$$\boldsymbol{\xi^{(u)}} = \sum_{j \in C^{(u)}} \frac{1}{|\mathscr{V}^{(j)}|} \boldsymbol{\theta^{(j)}}, \forall u \in \{1, \ldots, N_A\}. \quad (6)$$

These author-topic "barycenter" vectors $\{\boldsymbol{\xi^{(u)}}\}$ are each normalized to sum to 1 (Equation 7).

$$\bar{\boldsymbol{\xi}}^{\boldsymbol{(u)}} = \frac{1}{\sum_{k=1}^{KT} \xi_k^{(u)}} \boldsymbol{\xi^{(u)}}, \forall u \in \{1, \ldots, N_A\}. \quad (7)$$

3.3.2. Diffused Topic Distributions

Given distance metric $d(\cdot, \cdot) : \Delta^{\nu-1} \times \Delta^{\nu-1} \to \mathbb{R}_{\geq 0}$ a κ-nearest neighbors graph, $\mathcal{X}$, defined over the topic vectors $\{\boldsymbol{\phi^{(k)}} \in \Delta^{\nu-1}\}$, is formed. The vertices of $\mathcal{X}$ are the topic vectors. The edges in $\mathcal{X}$ are undirected, weighted by the distances between topic vector pairs, $d(\boldsymbol{\phi^{(i)}}, \boldsymbol{\phi^{(j)}})$. The Hellinger metric is used (Equation (2)). Justification is explored in Section 3.4 and references (Wang et al., 2021; Matsuzoe, 2010; Nielsen, 2020; Sun & Marchand-Maillet, 2014).

The ensemble learning and smoothing technique is driven by the assumption of a smooth topic manifold. Topics in adjacent time chunks are assumed to be reasonably similar, as justified in prior temporal topic models (Cui et al., 2011; Malik et al., 2013; Wang & McCallum, 2006; Blei & Lafferty, 2006; Wang et al., 2021). However, each document-topic distribution $\boldsymbol{\theta^{(j)}}$, is only possibly nonzero in up to $K^{(t)}$ positions, corresponding to the $K^{(t)}$ topics learned by the LDA model for C_t. To mitigate the sparsity of document-topic and author-topic distributions, a diffusion process redistributes probability mass across similar chunk topics. Algorithm 2 (Appendix F) details one-step message passing on the κ-nearest neighbors graph $\mathcal{X}$ using its weighted adjacency matrix, $\boldsymbol{X}$. The diffusion process outputs document-topic vectors, $\{\boldsymbol{\theta^{(j)}}\}$, and author-topic vectors, $\{\boldsymbol{\psi^{(u)}}\}$, both of which are over the entire corpus vocabulary.

3.4. Information-Geometric Manifold Learning

Manifold learning posits that data exist within a lower-dimensional structure. Information-geometric manifold learning captures the unique geometric properties of probability distributions, including topic, document-topic, and author-topic vectors. Typical manifold learning methods rely on capturing local geometric structures using κ-nearest neighbors graphs. Such methods are essential for visualizing topic trends like temporal trajectories and bifurcations (Moon et al., 2019; Wang et al., 2021).

The Hellinger metric (Equation 2) handles small probabilities and is compatible with methods based on L_2-distance, enabling the use of fast vector search tools. Topic vectors $\{\boldsymbol{\phi^{(k)}}\}$ are first transformed element-wise via square root to create modified vectors $\{\boldsymbol{\phi^{(k)'}}\}$. FAISS applies product quantization and creates a flat index for these modified vectors. This index is used to construct approximate nearest neighbors graphs for both $\mathcal{X}$ and $\mathcal{Y}$, reducing computational complexity as the number of authors and topics increase. The index and nearest neighbor graphs form the basis for hierarchical clustering and topic dendrogram construction. Furthermore, the approximate nearest neighbors graphs allow for meso-scale tuning of the manifold learning process, prior to applying the PHATE embedding algorithm.

3.5. Topic Dendrogram Construction

The MSTML topic dendrogram $\{\mathscr{D}; \{p_m\}\}$ is constructed in two stages. In the first, tree topology is determined using agglomerative hierarchical clustering of topic vectors $\{\boldsymbol{\phi^{(k)}}\}$. Second, probabilities $\{p_m\}$ are assigned to the internal nodes. The dendrogram provides a multi-scale representation of topic relationships and also links together the observed text and co-author network data.

For the topology of $\mathscr{D}$, LDA-derived topic vectors $\{\boldsymbol{\phi^{(k)}}\}$ are treated as multinomial distributions over the vocabulary $\mathcal{V}$. FAISS is used to create an approximate κ-nearest neighbors graph, $\mathcal{X}$, which captures local structures among the topic vectors. This graph is critical for agglomerative clustering, as it ensures that meso-scale relationships are preserved, aligning the clustering process with the diffusion-

based PHATE embeddings used later. The agglomerative clustering step iteratively merges topic clusters based on the Hellinger distance and chosen linkage method. Each merge produces an internal node m with a height h_m, representing a dissimilarity score between merged clusters. These heights h_m naturally increase up the tree but will exceed the Hellinger distance bounds of $[0, 1]$ when using Ward's linkage. For consistent interpretations with various linkages and distances, heights are re-normalized to the $[0, 1]$ interval.

To compute the internal node probabilities $\{p_m\}$, MSTML adopts a hierarchical random graph (HRG) model (Clauset et al., 2008), treating $\mathscr{E}_m$, L_m, and R_m as random variables based on author-topic embeddings $\{\boldsymbol{\psi^{(u)}}\}$. Authors are associated with chunk topics, which map directly to dendrogram leaves. Following Clauset et al. (2008), the MLE estimator for internal node probabilities is defined as $\hat{p}_m = \frac{\mathscr{E}_m}{L_m R_m}$, approximated here as $\hat{p}_m \approx \frac{\mathbb{E}[\mathscr{E}_m]}{\mathbb{E}[L_m]\mathbb{E}[R_m]}$ under the assumption of weak correlation between $\mathscr{E}_m$, L_m, and R_m.

This approach avoids the computational complexity of MCMC sampling by leveraging topic manifold clustering to learn the dendrogram topology only once. The dendrogram topology captures temporal topic trends and multi-scale author-topic relationships, forming the foundation for MSTML's probabilistic framework. As suggested by Clauset et al. (2008), the topology itself need only be "close" to optimal in order to estimate the internal node probabilities and evaluate likelihoods. Probabilities $\{p_m\}$ are computed as the ratio of expected edges over possible edges, conditioned on the LDA-learned topic distributions (Equation (8)). Appendix F.3 defines $\hat{\mathscr{E}}_m$, $\hat{L}_m$, and $\hat{R}_m$.

$$p_m \triangleq \frac{\hat{\mathscr{E}}_m}{\hat{L}_m \hat{R}_m}, \forall m \in \mathscr{D} \tag{8}$$

3.6. Multi-Scale Analysis

The MSTML dendrogram enables multi-scale analysis by leveraging hierarchical clustering and truncation. Each internal node m is assigned a height h_m, rescaled to $[0, 1]$ from the original Hellinger distances used during clustering. A cut height $h \in [0, 1]$ truncates the dendrogram, forming disjoint clusters of topics and authors. At height h, each cluster corresponds to a meta-topic formed by merging all leaf nodes in the subtree below any internal node m with $h_m \leq h$. Varying h is equivalent to horizontally slicing the tree in Figure 2 at different levels, producing topic partitions of differing granularities.

Truncating the dendrogram at height h allows analysts to study topics and author collaborations at multiple levels of abstraction. Fine-grained partitions at lower h values isolate niche topics, while higher h values merge topics into broader clusters. The likelihood of connections between authors in different clusters can be quantified using the probabilities $\{p_m\}$ defined in Equation (4), providing a systematic framework for exploring collaboration and topic diversity across scales.

3.7. Interdisciplinarity Scoring

Interdisciplinarity is assessed at both individual co-author links and document co-author lists.

First, each link $(u, v) \in \mathscr{E}$ in the co-author network $\mathscr{G}$ is assigned a link likelihood p_m, based on maximum likelihood estimates under a mixed-membership stochastic blockmodel derived from the topic dendrogram (Appendix G). These estimates assume that latent topic distributions drive collaborations, an assumption that holds better at higher levels of the dendrogram. The p_m link rankings identify unlikely candidate links rather than providing precise likelihood estimates.

Second, for each document j, interdisciplinarity is computed using topic distributions of the contributing authors. Each distribution $\boldsymbol{\psi^{(u)}}$ is modified to retain only the top N_{hot} values above a threshold τ, yielding $\boldsymbol{\psi^{(u)'}}$. A weighted group-level topic distribution $\boldsymbol{\Omega^{(j)}}$ is then calculated as the normalized weighted sum, $\boldsymbol{\Omega^{(j)}} = \frac{\sum_u w_u \cdot \boldsymbol{\psi^{(u)'}}}{\sum_u w_u}$, where $w_u = \sqrt{N_D^{(u)}}$ scales the contribution of each author. Finally, the interdisciplinarity score for j is obtained by multiplying the entropy of $\boldsymbol{\Omega^{(j)}}$ by the total author weight:

$$\text{Interdisciplinarity(j)} = H(\boldsymbol{\Omega^{(j)}}) \cdot \sum_u w_u.$$

This score prioritizes authors with more publications, reflecting stronger confidence in their topic expertise. It also highlights collaborations between authors who historically publish in consistent but distinct topics, emphasizing the notable nature of these interdisciplinary works.

4. EXPERIMENTS AND DISCUSSION

4.1. Data Preparation

Multiple document sets were used to validate the model. The body of this exposition highlights MSTML applied to a corpus sourced from arXiv, with additional case studies included in the appendix. arXiv includes manual category labels, which are user specified at the time of publication. Appendix C shows a distribution of the top category labels on arXiv. All arXiv data, including abstracts, author lists, and publication dates, were extracted from JSON-formatted source data from Kaggle. "arxiv-stat-ml" is a combined corpus of statistics and machine learning.

Code was written in Python and CPython. Using the pandas library, a dataframe was instantiated for each document set.

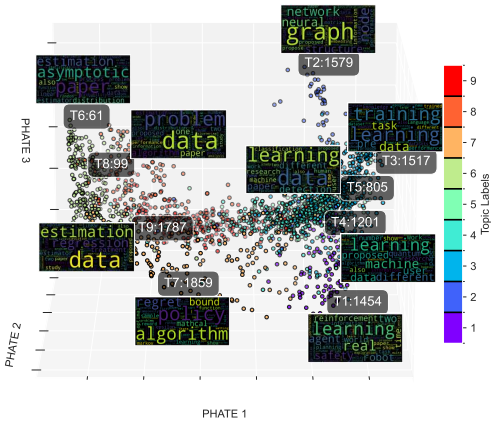


Figure 3. arxiv-stat-ml topic manifold: 9 meta topics at $h = 0.55$.

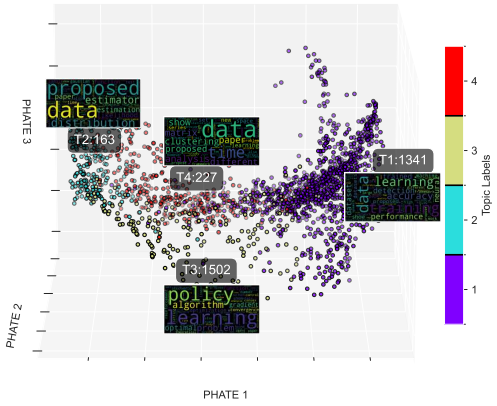


Figure 4. arxiv-stat-ml topic manifold: 4 meta topics at $h = 0.68$.

The dataframes were sorted by publication date. The natural language toolkit (NLTK) and gensim libraries provided lemmatization and stopword filtering, to create an initial vocabulary. Terms occurring in more than 99.5% of documents or in only 1 document were removed. All documents with more than 20 co-authors were also removed (Appendix J). An LDA model was then trained over the entire corpus for term-relevance filtering with $\lambda = 0.4$, taking the top 2000 words for each of 50 topics, resulting in a filtered vocabulary. Author lists associated with each row in the dataframe were disambiguated using fuzzy matching, and assigned unique, integer-valued author IDs.

4.2. Topic Manifold Exploration

Hellinger-PHATE embeddings of the topic manifold use $K^{(t)} = \max\left(\min\left(4, N_D^{(t)}\right), \frac{N_D^{(t)}}{100}\right)$ as the number of topics for time chunk t. This is based on reasonable ballpark estimates of the number of documents needed for LDA to sufficiently learn a distinct topic representation, which is also dependent on the vocabulary length and indeed, the language as well. The minimum of 4 topics (or $N_D^{(t)}$ topics in extreme cases), is used to encourage consistency in temporal topic alignment in cases where a specific time chunk may have a small number of publications. The Hellinger-PHATE manifold embeddings and hierarchical topic model enable unique multi-scale visualizations of the topic manifold. The arxiv-stat-ml manifold is represented as a saddle-like surface in three dimensions (Figures 3 and 4).

A key advantage of the smooth geometric alignment of the topic ensemble is that the manifold can be traversed from point to point to understand topic space and temporal relationships. Intermediary topics between coarse-grained concepts are revealed (Figure 5, Appendix A).

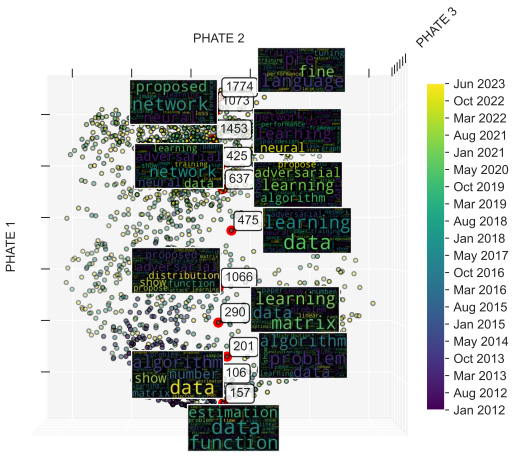


Figure 5. Traversing the topic manifold along PHATE1 separates statistics (bottom) and machine learning (top). The smooth geometry provides an intuitive understanding of intermediary, bridging topics. The selected points show a smooth transition between statistical estimation theory and fine-tuning of neural network language models.

MSTML takes manifold learning and hierarchical clustering a step further, by joining co-author network data with the smooth PHATE-Hellinger manifold. This allows one to understand trends for specific communities of authors, which can be divided by meta topics at any dendrogram cut height. Figures 6, 7 show example analyses for the graph learning and causal inference communities. These communities are defined by the cut-height, which partitions the co-author network. The network is also segmented based on time chunks, chosen appropriately based on the particular temporal characteristics of the corpus. Approximately 11 years of data were divided into 10 "mega chunks."

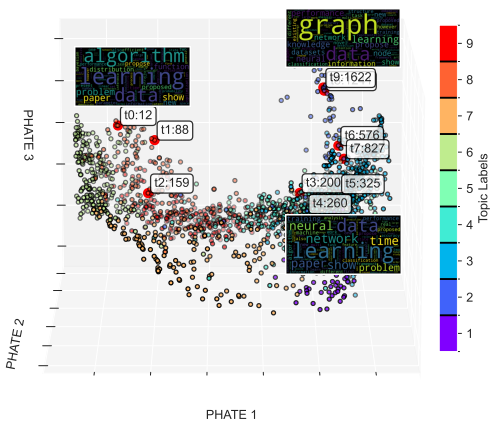


Figure 6. The graph learning topic (meta topic 2) forms an isolated branch, stemming from authors who started in statistics.

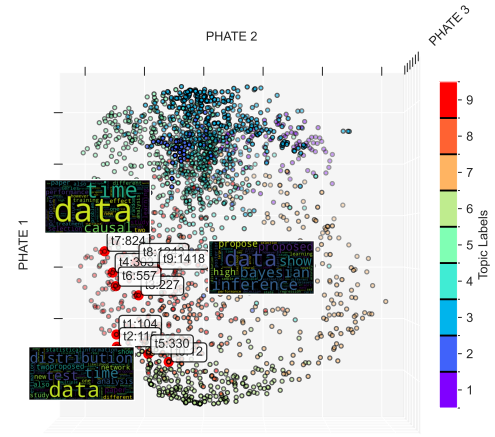


Figure 7. The causal inference community (meta topic 9) did not shift much over the last decade.

An additional critical advantage of the MSTML multi-modal and multi-scale approach lies in the generative probabilistic model imposed by the dendrogram. Recall that each internal node in $\mathscr{D}$ is parameterized by probability p_m. These p_m values are estimated according to independence assumptions which model each author as members of multiple heterogeneous topic communities. However, link independence is not a particularly good assumption, given that link prediction methods typically rely on correlation, using the number of common neighbors between two nodes to predict missing links (Lu & Zhou, 2011; Liben-Nowell & Kleinberg, 2003).

Despite this, there are also logical justifications for treating links as uncorrelated. Namely, while links are locally correlated, the clustering effect fades at meso and global scales. Due to network percolation, a massive connected component emerges when the mean node degree is greater than 1, which is easily satisfied by most co-author networks (Appendix J). An alternative justification is empirical. First, document indisciplinarity scores tend to be correlated with the number of different category labels that the associated author-sub-network published in, which are meta-data that were not used for model training (Appendix I). Second, link likelihood scores, $\{p_m\}$, tend to be correlated with the frequency of links between topic communities. This is true of the $\{p_m\}$ values by definition, but it is further empirically justified (Appendix G).

Interdisciplinarity pairwise topic scores can be ranked across the corpus, revealing the most critical bridging documents. Documents that score high in pairwise interdisciplinarity were authored by scientists with major contributions in at least two disciplines. Perhaps unsurprisingly, this score tends to favor survey articles (Table 1 and Appendix D).

Table 1. Highest Pairwise Interdisciplinarity Scores at $h = 0.68$.

Topic Pair	Title of Document with Highest Score
(1, 2)	Textbooks Are All You Need
(1, 3)	Regularization and Variance-Weighted Regression Achieves Minimax Optimality...
(1, 4)	Stein's Method Meets Computational Statistics: A Review of Some Recent Developments
(2, 3)	A Comprehensive Survey on Pretrained Foundation Models: ...BERT to ChatGPT
(2, 4)	Re-ViLM: Retrieval-Augmented Visual Language Model for Zero and Few-Shot...
(3, 4)	FedML: A Research Library and Benchmark for Federated Machine Learning

5. CONCLUSION

MSTML combines geometry-driven ensemble topic modeling with hierarchical random graph models, offering a novel framework for analyzing multi-disciplinary relationships in co-authored document corpora. MSTML embeds LDA-derived topics into a Hellinger-PHATE manifold and links topics to co-author networks to produce interpretable visualizations. The integration of temporal smoothing and diffusion processes ensures that both topic continuity and emerging trends are captured, while multi-scale dendrogram truncations allow for granular analyses of topic clusters and networks. Future work could integrate neural topic models, such as BERTopic, into the MSTML framework. Adaptive k-nearest neighbor graphs, which dynamically adjust to local topic density, could refine manifold construction and improve scalability. Automating additional key hyperparameter choices would enhance accessibility of the framework.

IMPACT STATEMENT

The societal impact of this work aligns with the broader implications typically associated with advancements in machine learning. As with any technological progress in this field, it is crucial to consider the potential ethical consequences of applying these advanced methods. For the techniques proposed here, concerns about ethical misuse are likely related to the extraction of data that users may consider private, even though the source information is publicly available. Modern data science and machine learning have proven to be much more effective at extracting insights than was previously anticipated, particularly in the digital age. While co-authorship data is public, this does not imply that authors are willing to have their collaboration and authorship behaviors analyzed in certain ways. On a more positive note, the MSTML method could facilitate more efficient and effective scientific collaborations in the future. It could help researchers identify meaningful connections and insights that are crucial for advancing their fields. Historically, scientific research has been linked to progress in reducing poverty and improving public health outcomes, driven by centuries of collaboration. In summary, while the methods proposed may enhance scientific collaboration, it is critical to ensure that their application benefits society without infringing on individual privacy or trust.

References

Adamic, L. A. and Adar, E. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

Beltagy, I., Lo, K., and Cohan, A. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3615–3620, 2019.

Blei, D. M. and Lafferty, J. D. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120, 2006.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022, 2003.

Burke, L., Pazdernik, K., Fortin, D., Wilson, B., Goychayev, R., and Mattingly, J. Nukelm: Pre-trained and fine-tuned language models for the nuclear and energy domains. *Pacific Northwest National Laboratory Technical Report*, 2023.

Clauset, A., Moore, C., and Newman, M. E. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

Cui, W. a. S. L., Tan, L., Shi, C., Song, Y., Gao, Z., Qu, H., and Tong, X. Textflow: Towards better understanding of evolving topics in text. In *IEEE transactions on visualization and computer graphics*, volume 17, pp. 2412–2421. IEEE, 2011.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

Gan, J. and Qi, Y. Selection of the optimal number of topics for lda topic model—taking patent policy analysis as an example. *Entropy*, 23(10):1301, 2021.

Hellinger, E. *Neue Begründung der Theorie der quadratischen Formen von unendlich vielen Veränderlichen*, volume 136. 1909.

Hofmann, T. Probabilistic latent semantic analysis. In *UAI*, volume 99, pp. 289–296, 1999.

Holland, P. W., Laskey, K. B., and Leinhardt, S. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

Hwang, S.-Y., Wei, C.-P., Lee, C.-H., and Chen, Y.-S. Coauthorship network-based literature recommendation with topic model. *Online Information Review*, 41(3):318–336, 2017.

Jegou, H., Douze, M., and Schmid, C. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2010.

Johnson, J., Douze, M., and Jegou, H. Billion-scale similarity search with gpus. In *IEEE Transactions on Big Data*, pp. 535–547, 2019.

Karrer, B. and Newman, M. E. J. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

Kenton, J. D. M.-W. C. and Toutanova, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. 1(2), 2019.

Krstovski, K., Smith, D. A., Wallach, H. M., and McGregor, A. Efficient nearest neighbor search in the probability simplex. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, pp. 101–108, 2013.

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

Kumar, A., Singh, S. S., Singh, K., and Biswas, B. Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications*, 553:124289, 2020.

Lee, D. D. and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755): 788–791, 1999.

Liben-Nowell, D. and Kleinberg, J. The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM)*, pp. 556–559, 2003.

Lu, K., Cai, X., Ajiferuke, I., and Wolfram, D. Vocabulary size and its effect on topic representation. *Information Processing & Management*, 53(3):653–665, 2017.

Lu, L. and Zhou, T. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

Lu, X. and Szymanski, B. K. A regularized stochastic block model for the robust community detection in complex networks. *Scientific Reports*, 9:13247, 2019.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., et al. Applying lda topic modeling in communication research: Toward a valid and reliable methodology. pp. 13–38, 2021.

Malik, S., Smith, A., Hawes, T., Papadatos, P., Li, J., Dunne, C., and Shneiderman, B. Topicflow: Visualizing topic alignment of twitter data over time. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 720–726. IEEE, 2013.

Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 978-0521865715.

Matsuzoe, H. Statistical manifolds and affine differential geometry. In *Probabilistic Approach to Geometry*, pp. 303–322, 2010.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, 2013.

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, S., Chen, W. S., Yim, X., van den Elzen, A., Hirn, K., Wolf, R., Krishnaswamy, D., Wolf, G., and Krishnaswamy, S. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37: 1482–1492, 2019.

Nassar, H., Benson, A. R., and Gleich, D. F. Pairwise link prediction. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 386–393, 2019.

Nielsen, F. An elementary introduction to information geometry. *Entropy*, 22(10):1100, 2020.

Okamura, K. Interdisciplinarity revisited: evidence for research impact and dynamism. *Palgrave Communications*, 5(1), 2019.

Pearson, K. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11): 559–572, 1901.

Roweis, S. T. and Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (5500):2323–2326, 2000.

Sievert, C. and Shirley, K. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70, Baltimore, Maryland, USA, 2014. Association for Computational Linguistics.

Sparck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

Sun, K. and Marchand-Maillet, S. An information geometry of statistical manifold learning. pp. 1–9, 2014.

Ullah, M., Shahid, A., Din, I. U., Roman, M., Assam, M., Fayaz, M., Ghadi, Y., and Aljuaid, H. Analyzing interdisciplinary research using co-authorship networks. *Complexity*, 2022(1):2524491, 2022.

van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Wang, X. and McCallum, A. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 424–433. ACM, 2006.

Wang, Y., Hougen, C., Oselio, B., Dempsey, W., and Hero, A. A Geometry-Driven Longitudinal Topic Model. *Harvard Data Science Review*, 3(2), jun 30 2021. https://hdsr.mitpress.mit.edu/pub/0v7qw6jf.

Zhang, Y., Wu, M., Zhang, G., and Lu, J. Stepping beyond your comfort zone: Diffusion-based network analytics for knowledge trajectory recommendation. *Journal of the Association for Information Science and Technology*, 74 (7):775–790, 2023.

Zhu, Y., Quan, L., Chen, P., Kim, M. C., and Che, C. Predicting co-authorship using bibliographic network embedding. In *Journal of the Association for Information Science and Technology*, pp. 388–401, 2023.

A. arxiv-stat-ml Manifold Traversal Paths

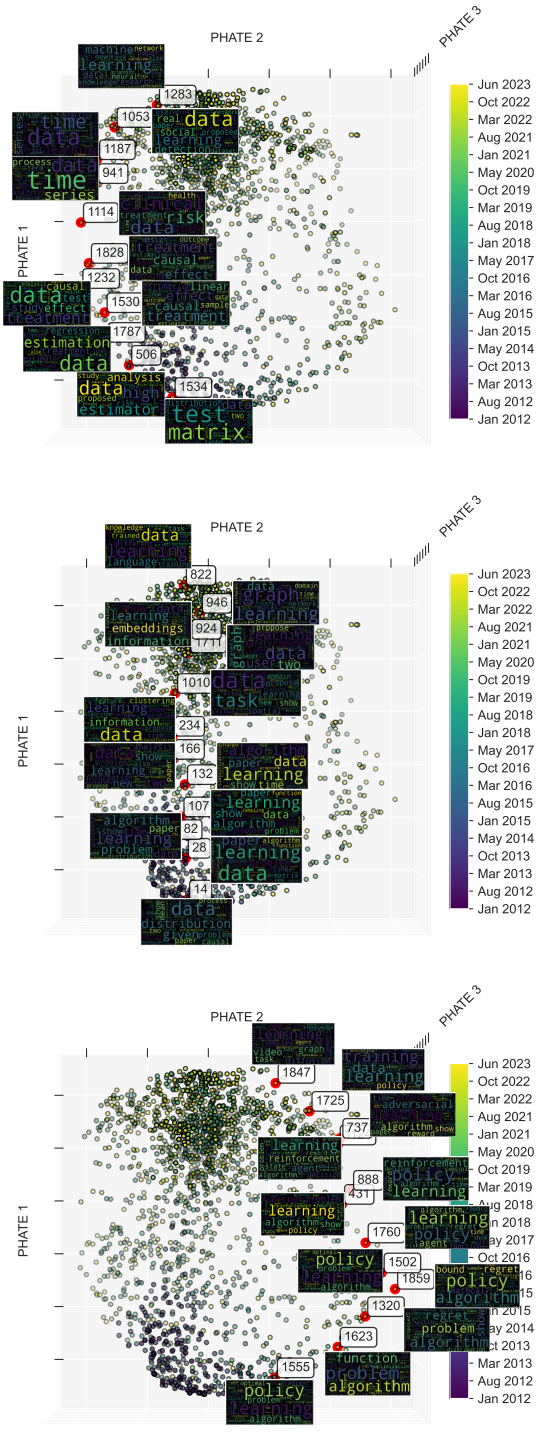
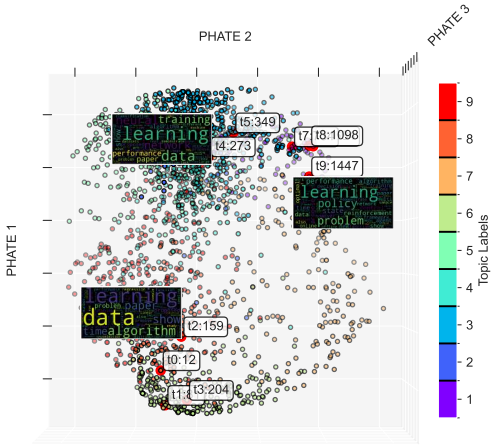


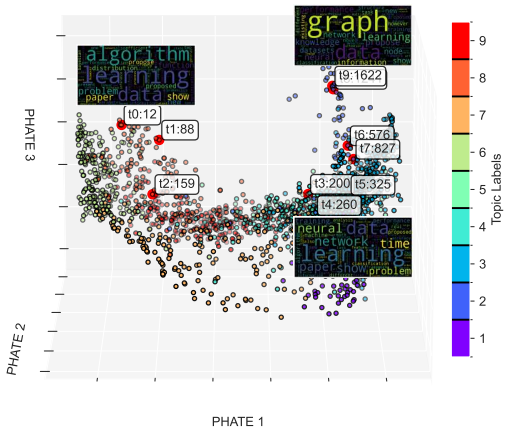
Figure 8. Sampled traversal paths along PHATE1 reveal topic bridges between statistics and machine learning. (Top) Bayesian statistics, linear regression, healthcare, and time series; (middle) sampling, clustering, and embeddings; (bottom) regret, optimization, and reinforcement learning. Traversal paths show intuitve topic relationships, enabled by the topic geometry.

B. Author Community Trends

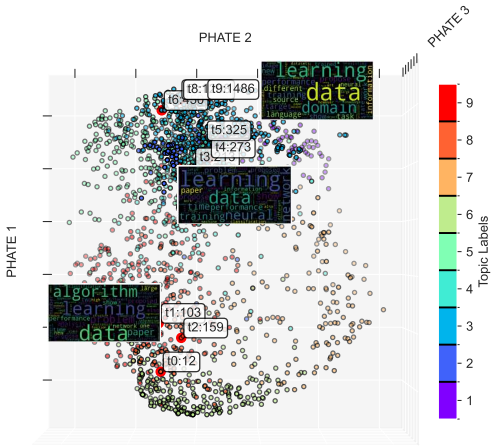
There are multiple options for partitioning the co-author network into communities. Unlike the standard network science techniques, MSTML uses the topic and chunk index information to partition the co-author network. In this appendix, each author is mapped to the meta topic cluster that is of greatest probability mass in the author distributions at the particular cut level of the dendrogram. Here, the cut height is $h = 0.55$, which results in 9 clusters. Chunks of documents correspond to 1 month time periods for the arxiv-stat-ml data. We group these chunks into mega chunks, which are adjacent intervals of chunks, denoted by the times $t0$ through $t9$ in each of the figures. These diagrams show topic shift over time of each particular community. For example, the reinforcement learning topic community (meta topic 1), did not primarily publish in reinforcement learning across their community early in the corpus. In fact many communities published more in the statistics-related topics toward the bottom of the PHATE diagrams, even if they later published in topics related to machine learning, toward the top of the diagrams. These diagrams give an idea of how each community shifted across the topic manifold over the last decade, since 2012. The meta topics are hand-labeled by inspecting not only these wordclouds but also by looking at the wordclouds of other explorations of the manifold, such as those displayed in Appendix A and Appendix E.



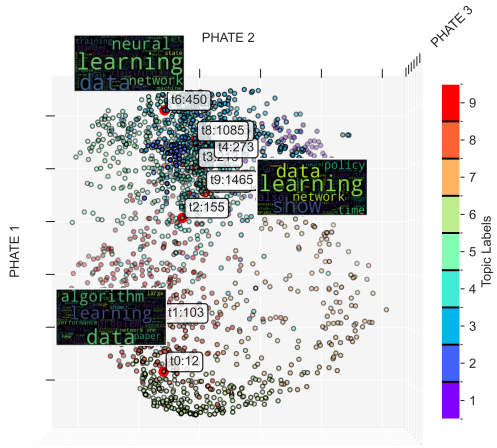
(a) Meta topic 1: reinforcement learning and robotics



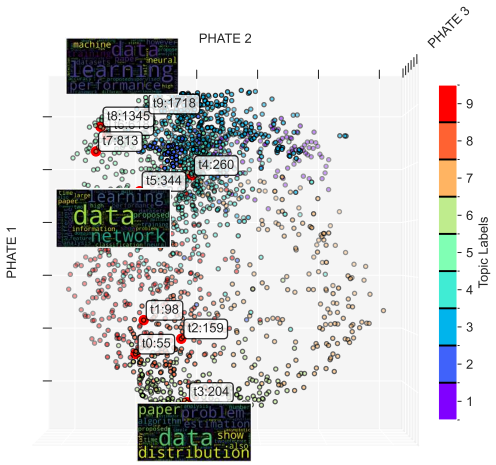
(b) Meta topic 2: graph learning



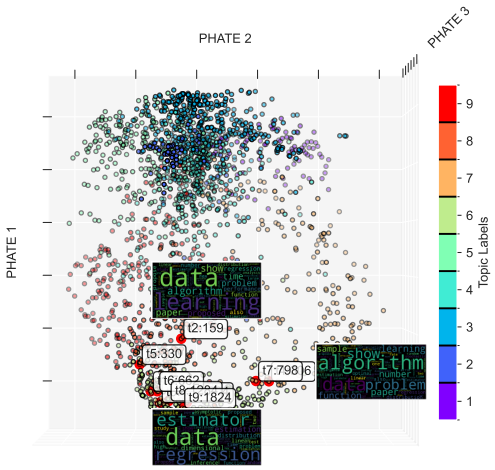
(c) Meta topic 3: language models



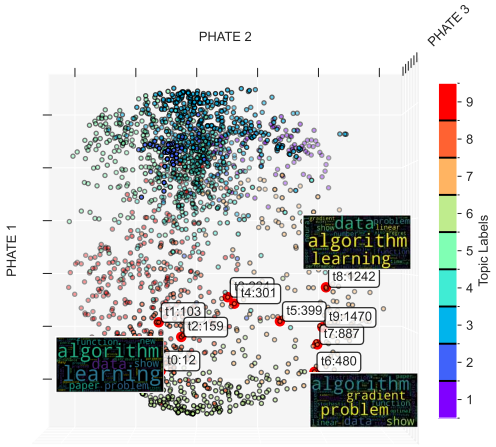
(d) Meta topic 4: security and protein models



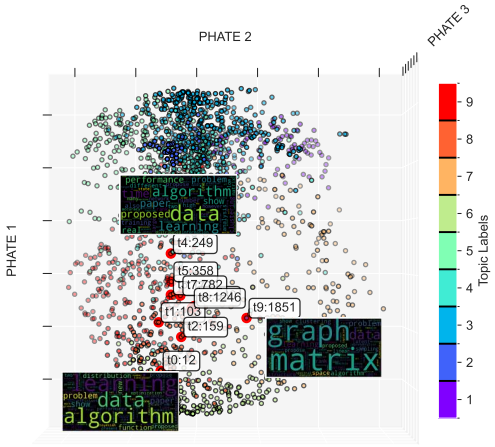
(a) Meta topic 5: supervised learning



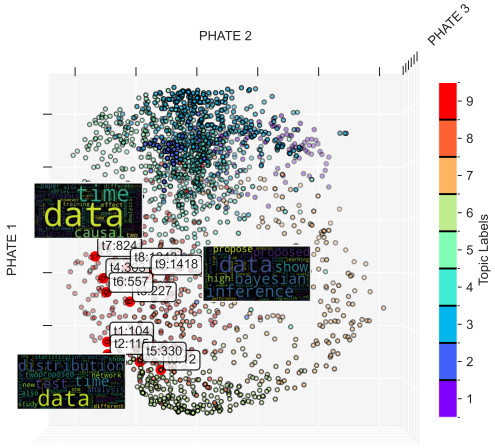
(b) Meta topic 6: statistical estimation theory



(c) Meta topic 7: decision theory and optimization



(d) Meta topic 8: data science mixed



(e) Meta topic 9: causal inference

C. arXiv Category Label Distribution

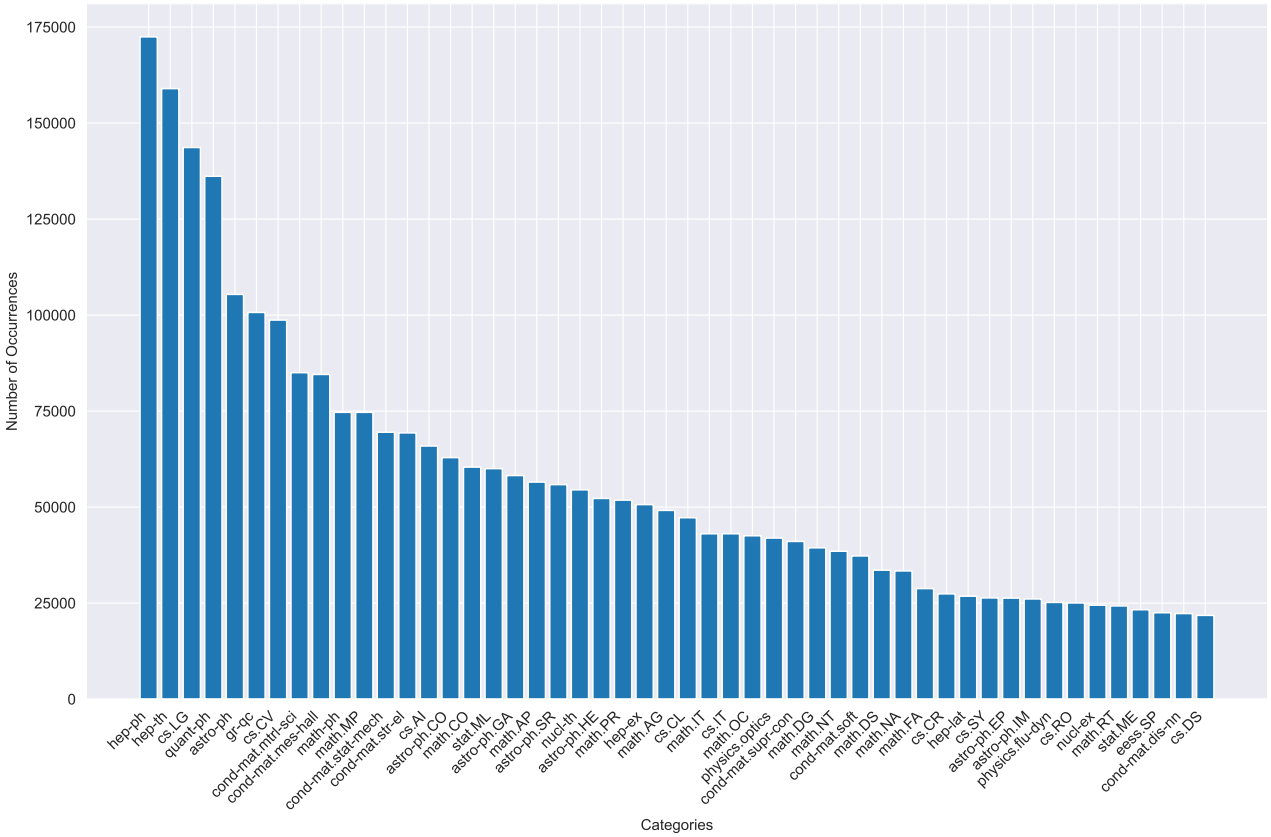


Figure 11. The frequency of documents of each category label in arXiv, between the years 2012 and 2023. Notably, high-energy physics and related topics are the most popular in arXiv, with machine learning topics following close behind.

D. Topic Pair Interdisciplinarity Scores

D.1. arxiv-stat-ml at $h = 0.55$ (9 Meta Topics)

Table 2. Highest Pairwise Interdisciplinarity Scores at $h = 0.55$.

Topic Pair	Title	First Author	Score
(1, 2)	FedML: A Research Library and Benchmark for Federated ML	He, C.	116.62
(1, 3)	Language to Rewards for Robotic Skill Synthesis	Yu, W	74.58
(1, 4)	From ML to Robotics: Challenges for Embodied Intelligence	Roy, N.	65.73
(1, 5)	Extending the WILDS Benchmark for Unsupervised Adaptation	Sagawa, S.	104.62
(1, 6)	Survey on Pretrained Foundation Models: BERT to ChatGPT	Zhou, C.	117.69
(1, 7)	Survey on Pretrained Foundation Models: BERT to ChatGPT	Zhou, C.	116.86
(1, 8)	FedML: A Research Library and Benchmark for Federated ML	He, C.	116.70
(1, 9)	Survey on Pretrained Foundation Models: BERT to ChatGPT	Zhou, C.	117.11
(2, 3)	DIG: A Library for Graph Deep Learning Research	Liu, M.	56.27
(2, 4)	Task-Aware Effective Brain Connectivity for fMRI with GNNs	Yu, Y.	62.49
(2, 5)	Trustworthy Graph Learning: Reliability, Explainability, Privacy	Wu, B.	105.06
(2, 6)	Survey on Pretrained Foundation Models BERT to ChatGPT	Zhou, C.	116.82
(2, 7)	Survey on Pretrained Foundation Models: BERT to ChatGPT	Zhou, C.	114.00
(2, 8)	Survey on Pretrained Foundation Models: BERT to ChatGPT	Zhou, C.	117.65
(2, 9)	Survey on Pretrained Foundation Models: BERT to ChatGPT	Zhou, C.	117.37
(3, 4)	FedML: A Research Library and Benchmark for Federated ML	He, C.	115.98
(3, 5)	Discussions on Semi-Automatic Approximate Bayesian	Andrieu, C.	81.25
(3, 6)	Textbooks Are All You Need	Gunasekar, S.	82.70
(3, 7)	Textbooks Are All You Need	Gunasekar, S.	82.70
(3, 8)	EmoNets: Multimodal Learning for Emotion Recognition	Kahou, S. E.	89.59
(3, 9)	Understanding Self-Predictive Learning for RL	Tang, Y.	69.76
(4, 5)	Summary of the ComParE COVID-19 Challenges	Coppock, H.	61.38
(4, 6)	Textbooks Are All You Need	Gunasekar, S.	84.26
(4, 7)	Textbooks Are All You Need	Gunasekar, S.	82.44
(4, 8)	Discussions on Semi-Automatic Approximate Bayesian	Andrieu, C.	67.60
(4, 9)	Discussions on Semi-Automatic Approximate Bayesian	Andrieu, C.	74.26
(5, 6)	FedML: A Research Library and Benchmark for Federated ML	He, C.	104.31
(5, 7)	FedML: A Research Library and Benchmark for Federated ML	He, C.	110.04
(5, 8)	RL Unplugged: Benchmarks for Offline Reinforcement Learning	Gulcehre, C.	99.66
(5, 9)	RL Unplugged: Benchmarks for Offline Reinforcement Learning	Gulcehre, C.	99.72
(6, 7)	Survey on Pretrained Foundation Models: BERT to ChatGPT	Zhou, C.	111.01
(6, 8)	Survey on Pretrained Foundation Models: BERT to ChatGPT	Zhou, C.	117.00
(6, 9)	FedML: A Research Library and Benchmark for Federated ML	He, C.	113.01
(7, 8)	Survey on Pretrained Foundation Models: BERT to ChatGPT	Zhou, C.	117.69
(7, 9)	Survey on Pretrained Foundation Models: BERT to ChatGPT	Zhou, C.	112.68
(8, 9)	Survey on Pretrained Foundation Models: BERT to ChatGPT	Zhou, C.	112.79

D.2. arxiv-stat-ml at $h = 0.68$ (4 Meta Topics)

Table 3. Highest Pairwise Interdisciplinarity Scores at $h = 0.68$.

Topic Pair	Title	First Author	Score
(1, 2)	Textbooks Are All You Need	Gunasekar, S.	75.59
(1, 3)	Regularization and Variance-Weighted Regression Achieves Minimax…	Kitamura, T.	64.77
(1, 4)	Stein's Method Meets Computational Statistics: A Review of Some…	Anastasiou, A.	75.46
(2, 3)	A Comprehensive Survey on Pretrained Foundation Models: BERT to ChatGPT	Zhou, C.	111.01
(2, 4)	Re-ViLM: Retrieval-Augmented Visual Language Model for Zero and…	Yang, Z.	87.69
(3, 4)	FedML: A Research Library and Benchmark for Federated Machine Learning	He, C.	106.66

E. Term Relevance Hyperparameter λ

Term relevance (Equation (1)) was formally defined in (Sievert & Shirley, 2014) and repeated here for convenience: $r(w, k \mid \lambda) = \lambda \log P(w \mid k) + (1 - \lambda) \log \left(\frac{P(w|k)}{P(w)} \right)$. Term relevance is uniquely useful in the MSTML model because of the ensemble learning approach. First, term relevance is used for initial filtering and reduction of the corpus vocabulary. Second, it can be used as an adjustable parameter in conjunction with the Hellinger-PHATE embedding to reveal constrastive properties of particular chunk topics. The grid of wordclouds is organized from meta topic 1 on the left to meta topic 9 on the right, based on $h = 0.55$ in the arxiv-stat-ml data.

Figure 12. Grid of the 9 chunk topics chosen as representative points (columns), with each row corresponding to a different term-relevance hyperparameter ($\lambda \in \{1.0, 0.8, 0.6, \ldots, 0.0\}$ from top to bottom).

F. MSTML Method Details

F.1. MSTML Algorithm

Algorithm 1 is the main algorithm for training the MSTML model and learning the topic manifold and parameterized dendrogram. The primary steps of the algorithm are summarized - details of Steps 1 through 6 are included in Section 3.

Algorithm 1 Multi-Scale Topic Manifold Learning

Input: Document corpus C of size N_D; N_A associated authors in author (vertex) set, $\mathscr{V}$.
Input: Corpus chunk length in months, $M \in \mathbb{N}$; number of documents per topic, $N_\delta \in \mathbb{N}$; chunk-smoothing parameter $\gamma \in (0, 1)$, minimum term count threshold, $\zeta \in \mathbb{N}$, and maximum term frequency threshold $\epsilon \in (0, 1)$.
Output: A topic dendrogram parameterized by internal node probabilities $\{\mathscr{D}; \{p_m\}\}$.
Output: Author-topic vectors $\{\boldsymbol{\psi^{(u)}}\}_{u=1}^{N_A}$ with $\boldsymbol{\psi^{(u)}} \in \Delta^{K-1}$, $\forall u \in \mathscr{V}$, where K is the total number of chunk topics in the ensemble.
Output: A filtered term-vocabulary $\mathcal{V}$, of size $\nu = |\mathcal{V}|$.
Output: Topic word-frequency vectors $\{\boldsymbol{\phi^{(k)}}\}_{k=1}^{K}$ with $\boldsymbol{\phi^{(k)}} \in \Delta^{\nu-1}$, $\forall k$.
Step 1: Vocabulary filtering
Step 2: Corpus-chunking and smoothing
Step 3: Co-author network construction
Step 4: Topic model ensemble training
Step 5: Author embedding process
Step 6: Topic dendrogram construction
return $\{\mathscr{D}; \{p_m\}\}$, $\{\boldsymbol{\psi^{(u)}}\}_{u=1}^{N_A}$, $\{\boldsymbol{\phi^{(k)}}\}_{k=1}^{K}$, and $\mathcal{V}$.

F.2. Author Embedding Diffusion Algorithm

This one-step message-passing algorithm is used to diffuse probability mass among nearby chunk-topics. This is used to smooth the LDA topics and the author-topic distributions across chunks.

Algorithm 2 Author Embedding Diffusion

Input: $K \in \mathbb{N}$, $\nu \in \mathbb{N}$, and $\omega \in [0, 1]$.
Input: $\{\bar{\boldsymbol{\xi}}^{(\boldsymbol{u})} \in \Delta^{K-1}\}$, $\mathcal{X}$ and $\boldsymbol{X}$.
Output: $\{\boldsymbol{\psi^{(u)}} \in \Delta^{K-1}\}$
for each author $u \in \mathscr{V}$ **do**
  $\boldsymbol{q} \leftarrow \bar{\boldsymbol{\xi}}^{(\boldsymbol{u})}$
  for each topic $k \in \mathcal{X}$ **do**
    if $\bar{\xi}_k^{(u)} > 0$ **then**
      $q_k \leftarrow \bar{\boldsymbol{\xi}}_k^{(\boldsymbol{u})}$
    else
      Let $\mathcal{N}(k)$ be the neighbors of topic k in $\mathcal{X}$.
      if $\mathcal{N}(k)$ is not empty **then**
        $\boldsymbol{w} \leftarrow 1 - \frac{1}{\sum \boldsymbol{x_{k,\mathcal{N}(k)}}} \boldsymbol{x_{k,\mathcal{N}(k)}}$
        $\boldsymbol{\eta} \leftarrow \bar{\boldsymbol{\xi}}^{(\boldsymbol{u})}_{\mathcal{N}(k)}$
        $q_k \leftarrow q_k + \omega \boldsymbol{w}^T \boldsymbol{\eta}$
      end if
    end if
  end for
  $\boldsymbol{\psi^{(u)}} \leftarrow \frac{1}{\sum(\boldsymbol{q})} \boldsymbol{q}$
end for
return $\{\boldsymbol{\psi^{(u)}} \in \Delta^{K-1}\}$.

F.3. Link Likelihood Definitions

The variables from Equation (8) are defined as follows. U_l is the joint probability that author u belongs to the left sub-tree and author v belongs to the left sub-tree of the dendrogram. V_l is defined similarly. $\hat{\mathscr{E}}_m$ is the expected number of links between the left and right sub-trees of internal node m of the dendrogram. $\hat{L}_m$ and $\hat{R}_m$ are the expected numbers of authors that are members of topics in the left or right sub-trees of m, respectively.

$$\hat{\mathscr{E}}_m \triangleq \sum_{(u,v)\in\mathscr{E}} (U_l + V_l) \tag{9}$$

$$U_l \triangleq \psi_l^{(u)}(1-\psi_r^{(u)})\psi_r^{(v)}(1-\psi_l^{(v)}) \tag{10}$$

$$V_l \triangleq \psi_l^{(v)}(1-\psi_r^{(v)})\psi_r^{(u)}(1-\psi_l^{(u)}) \tag{11}$$

$$\hat{L}_m \triangleq \left(\sum_{u\in\mathscr{V}} \psi_l^{(u)}(1-\psi_r^{(u)})\right) \tag{12}$$

$$\hat{R}_m \triangleq \left(\sum_{v\in\mathscr{V}} \psi_r^{(v)}(1-\psi_l^{(v)})\right) \tag{13}$$

F.4. Document Interdisciplinarity Definitions

We compute the interdisciplinarity of a document j based on the topic distributions of the authors who contributed to it. Let $\boldsymbol{\psi^{(u)}}$ denote the topic distribution vector for author u, and let $\boldsymbol{\Omega^{(j)}}$ denote the weighted group-level topic distribution vector for document j. The computation proceeds as follows:

1. Weighted Group-Level Topic Distribution

The group-level topic distribution $\boldsymbol{\Omega^{(j)}}$ is computed as the weighted sum of thresholded author distributions, normalized by the total weight:

$$\boldsymbol{\Omega^{(j)}} = \frac{\sum_u w_u \cdot \boldsymbol{\psi^{(u)\prime}}}{\sum_u w_u} \tag{14}$$

where:

- $\boldsymbol{\psi^{(u)}}$: Original topic distribution vector for author u,
- $\boldsymbol{\psi^{(u)\prime}}$: Thresholded version of $\boldsymbol{\psi^{(u)}}$, defined below,
- $w_u = \sqrt{N_D^{(u)}}$: Weight for author u, proportional to the square root of the number of documents authored by u, $N_D^{(u)}$.

2. Thresholding of Author Distributions

Each author's topic distribution $\boldsymbol{\psi^{(u)}}$ is thresholded to retain only the top N_{hot} values above a threshold τ. The thresholded distribution $\boldsymbol{\psi^{(u)\prime}}$ is defined as:

$$\psi_t^{(u)\prime} = \begin{cases} \psi_t^{(u)}, & \text{if } \psi_t^{(u)} \geq \tau \text{ and } t \text{ is in the top } N_{\text{hot}} \text{ values of } \boldsymbol{\psi^{(u)}} \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

3. Entropy of $\boldsymbol{\Omega^{(j)}}$

The entropy of the group-level distribution $\boldsymbol{\Omega^{(j)}}$ measures the diversity of topics within the document:

$$H(\boldsymbol{\Omega^{(j)}}) = -\sum_t \Omega_t^{(j)} \log \Omega_t^{(j)} \tag{16}$$

where $\Omega_t^{(j)}$ is the t-th element of the vector $\boldsymbol{\Omega^{(j)}}$.

4. Final Interdisciplinarity Score

The interdisciplinarity score for document j incorporates both the entropy $H(\mathbf{\Omega}^{(j)})$ and the total weight of the contributing authors:

$$\text{Interdisciplinarity Score for } j = H(\mathbf{\Omega}^{(j)}) \cdot \sum_{u} w_u \tag{17}$$

G. Generative Model and Verification of Link Likelihood Estimation

The generative model underpinning MSTML is based on the ideas from hierarchical random graphs (HRG) (Clauset et al., 2008). In the HRG model, the hierarchical dendrogram represents a generative process known as a stochastic blockmodel (SBM) (Holland et al., 1983) with heterogeneous intra-community and inter-community link probabilities. In MSTML, this is taken a step further, as we allow authors to be members of multiple communities simultaneously, according to their topic distributions. In network science, this is known as the mixed membership stochastic blockmodel, or MMSBM, first pioneered in (Airoldi et al., 2008).

Statistically, it is an ill-posed challenge to estimate the true link likelihoods in most real-world scenarios. Co-author networks are sparse, meaning that the generative probabilities are small, and this leads to noisy observations that may not be reflective of the underlying affinities between various topic communities. Furthermore, there are inherent challenges with estimating the link likelihoods due to dependence on the learned topic distributions, which introduce additional noise and uncertainty. The learned author-topic distributions are data and hyperparameter dependent. Despite these concerns, the link likelihood estimates can still be useful if the link rankings are capable of identifying classes of unlikely links. In particular, the multi-scale MSTML framework enables the tuning of link rankings according to the characteristics of the data. After flagging potentially noteworthy "bridging" links between topics that are far away in the topic space, the strengths of the interpretable topic manifold framework are leveraged to further reveal why those particular links may have been scored as particularly unlikely.

G.1. Generative Model for the Co-Author Network

Using a mixed membership stochastic block model (MMSBM), authors are modeled as members of latent topic communities. This generative model ensures that authors can simultaneously belong to multiple communities with varying degrees of membership, which mirrors the LDA generative model for documents.

Author-Topic Distributions: Each author u is associated with a multinomial distribution $\boldsymbol{\psi^{(u)}}$ over K latent topics. The author-topic distributions are determined in practice by the LDA learning process, or by the ensemble methods in MSTML which include smoothing. For the purposes of simulation, author distributions are sampled from a symmetric Dirichlet prior:

$$\boldsymbol{\psi^{(u)}} \sim \text{Dirichlet}(\alpha)$$

where α is the concentration parameter controlling the extent of topic mixing.

Topic Affinity Matrix: An affinity matrix $\Pi \in \mathbb{R}^{K\times K}$ encodes the probabilities of connections between authors in K different topics. The diagonal elements Π_{ii} represent the likelihood of links within the same topic, while the off-diagonal elements Π_{ij} $(i \neq j)$ represent the likelihood of links between different topics. In the context of the multi-scale topic dendrogram, $\mathscr{D}$, we can think of Π as a particular instance, $\Pi^{(h)}$, associated with cut height h. Each p_m value for the leaf nodes at cut height h is then simply the maximum likelihood estimator of the true link likelihood, $\Pi_{ij}^{(h)}$, conditioned on the author-topic distributions.

Link Probabilities: For a specific pair of authors u and v, their link probability is:

$$P((u,v)\in\mathscr{E}) = \boldsymbol{\psi^{(u)}}\cdot\Pi\cdot\boldsymbol{\psi^{(v)}} \tag{18}$$

This equation accounts for the mixed membership of both authors and the underlying topic affinities encoded by Π.

Network Generation: For each author pair (u,v), an edge is sampled as a Bernoulli random variable with probability $P((u,v)\in\mathscr{E})$. The resulting graph $\mathscr{G}$ is a simple, undirected co-author network, which is enforced by ensuring that $u \neq v$ at each sampling step. Algorithm 3 describes the details of the generative model for the network.

G.2. Verifying p_m for Identifying Interdisciplinary Links

Given a simulated co-author network $\mathscr{G}$, we aim to estimate link likelihoods using maximum likelihood estimates of the probabilities of topic pair interactions. The definitions of $\hat{\mathscr{E}}_m$, $\hat{L}_m$ and $\hat{R}_m$ follow those in Appendix F.3. Then, the MLE

link probability is repeated here, as in Equation (8):

$$p_m \triangleq \frac{\hat{\mathscr{E}}_m}{\hat{L}_m \hat{R}_m}$$

This ratio reflects the observed number of edges normalized by the expected number of possible edges between topics L_m and R_m.

By simulating multiple co-author networks according to the generative model, we may gain confidence that p_m values are reasonable empirical estimators for the true, unobserved affinities defined by the Π matrix. The intent is to show that p_m, and derivatives of p_m, are capable of identifying unlikely links according to the underlying topic distributions. Assuming that the topic models and author-topic distributions are reliable, the p_m values could then be used to identify links that are particularly noteworthy for bridging authors from distant topic communities.

Ground Truth Rankings. The ground truth probability p_{uv} of a link between two authors u and v is computed directly as:

$$P((u,v) \in \mathscr{E}) = p_{uv} = \boldsymbol{\psi}^{(\boldsymbol{u})} \cdot \Pi \cdot \boldsymbol{\psi}^{(\boldsymbol{v})}.$$

MLE Rankings. The MLE-based ranking uses p_m values derived from observed edge counts, serving as an approximation to the true link probabilities:

$$P((u,v) \in \mathscr{E}) \approx \hat{p}_{uv} = \boldsymbol{\psi}^{(\boldsymbol{u})} \cdot \hat{\Pi} \cdot \boldsymbol{\psi}^{(\boldsymbol{v})},$$

where $\hat{\Pi}$ is the maximum likelihood estimator of the Π matrix formed by populating the entries with the relevant p_m values.

Kendall's Tau for Rank Correlation. To quantify the agreement between ground truth and MLE rankings, we compute Kendall's Tau rank correlation coefficient over multiple simulation trials. A high tau value indicates that p_m rankings are a reliable proxy for identifying links with low probabilities, even under sparse observations.

Sensitivity to Model Parameters The accuracy of p_m as an estimator of $\Pi_{ij}^{(h)}$ is influenced by several factors:

- **Number of Topics (K):** A higher K increases the complexity of the topic structure, potentially reducing the overlap between topics and increasing sparsity in observed edges. The K value is dependent on hyperparameters of the ensemble model, as well as the cut-height h.
- **Affinity Matrix (Π):** The relative magnitudes of diagonal (Π_{ii}) and off-diagonal (Π_{ij}) elements determine the balance between within-topic and cross-topic connections. Higher Π values improve p_m-based estimates due to a higher signal to noise ratio, which means fewer observed links necessary.
- **Sparsity:** Related to the above, sparse networks with low average degree introduce noise into $\hat{\mathscr{E}}_m$. p_m is generally less reliable for rare interactions.

Table 4 shows Kendall's Tau rank correlation between the ground truth least likely links, and the least likely links according to MSTML. Only the $\min(500, \lfloor \frac{|\mathscr{E}|}{20} \rfloor)$ most unlikely links are considered when comparing the two ranked lists. The table also shows the rank correlation scores when both the estimated p_m values and the ground truth probabilities are weighted by $w = \max(10, \sqrt{\deg(u)\deg(v)})$, which biases the rankings toward authors for which there are sufficient observations. The point of this experiment is to justify the definition of p_m, which depends on replacing $\mathscr{E}_m$, L_m, and R_m by their expectations, taken over possible author topic assignments. The table demonstrates empirical evidence that this assumption is reasonable, as demonstrated by the somewhat positive correlation between link rankings and ground truth ranked probabilities for small values of K, the number of topics. These experiments are still preliminary and would need to be investigated further to draw any stronger conclusions, however.

Table 4. Kendall's Tau Rank Correlation Across Simulation Trials for Different Parameters.

Avg Degree	Num Topics	Unweighted Tau (Mean)	Unweighted Tau (Variance)	Weighted Tau (Mean)	Weighted Tau (Variance)
5	3	0.229	0.001	0.241	0.002
5	5	0.230	0.001	0.250	0.001
5	10	0.221	0.002	0.217	0.002
10	3	0.230	0.001	0.245	0.001
10	5	0.231	0.000	0.233	0.001
10	10	0.249	0.001	0.220	0.001
15	3	0.233	0.000	0.241	0.000
15	5	0.228	0.001	0.232	0.000
15	10	0.224	0.001	0.237	0.000

Algorithm 3 Generative MMSBM for Co-Author Network

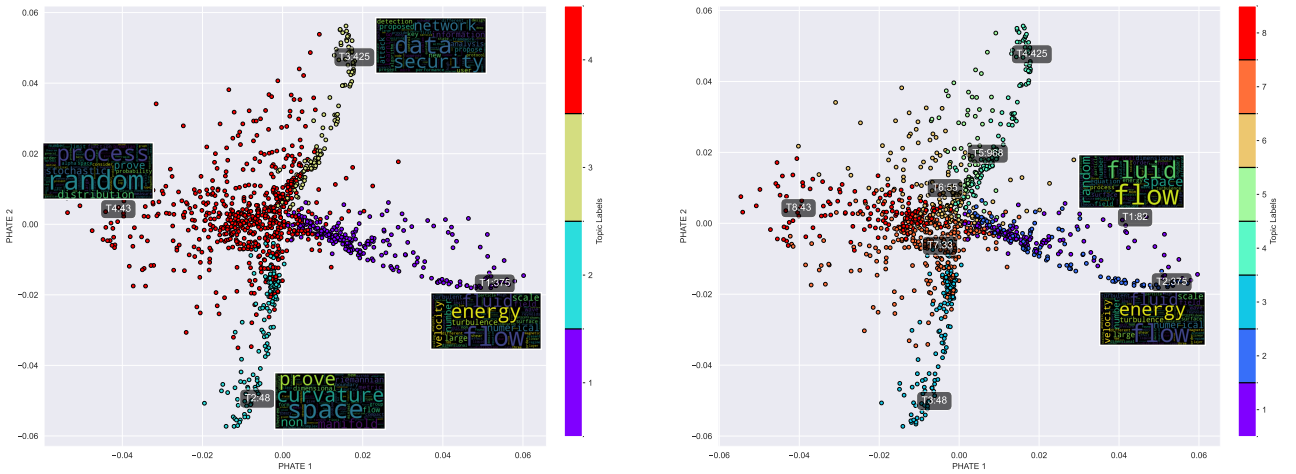
Input: Author-topic distributions $\boldsymbol{\psi^{(u)}} \sim \text{Dirichlet}(\alpha)$, $\forall u \in \{1, \ldots, N_A\}$
Input: $N_A, K \in \mathbb{N}$; $B \in \mathbb{N}$; $\alpha \in \mathbb{R}^+$; Inter-topic link affinity matrix $\Pi \in \mathbb{R}^{K \times K}$
Output: Co-author network $\mathscr{G}$
Initialize graph $\mathscr{G}$ with N_A nodes and zero edges
Set $b \leftarrow 0$ (current link count)
Step 1: Precompute Pairwise Link Probabilities
Compute matrix $\mathbf{P} \in \mathbb{R}^{N_A \times N_A}$, where:

$$P_{uv} = \boldsymbol{\psi^{(u)}} \cdot \Pi \cdot \boldsymbol{\psi^{(v)}} \quad \forall u, v \in \{1, \ldots, N_A\}, u \neq v$$

Set diagonal entries $P_{uu} \leftarrow 0$ to avoid self-loops.
Step 2: Ensure Every Author Has at Least One Edge
for $u \in \{1, \ldots, N_A\}$ **do**
  if $\deg(u)$ in $\mathscr{G}$ is 0 **then**
    Sample $v \sim \text{Multinomial}(\mathbf{P}[u,:]/\|\mathbf{P}[u,:]\|_1)$
    Add edge (u, v) to $\mathscr{G}$
    $b \leftarrow b + 1$
  end if
end for
Step 3: Sample Additional Links to Achieve Target Edge Count
Compute total target edges $B_{\text{target}} \leftarrow B$
while $b < B_{\text{target}}$ **do**
  Sample (u, v) from the upper triangle of $\mathbf{P}$ with probabilities proportional to P_{uv}
  if $(u, v) \notin \mathscr{G}$ **then**
    Add edge (u, v) to $\mathscr{G}$
    $b \leftarrow b + 1$
  end if
end while
Output:
return $\mathscr{G}$

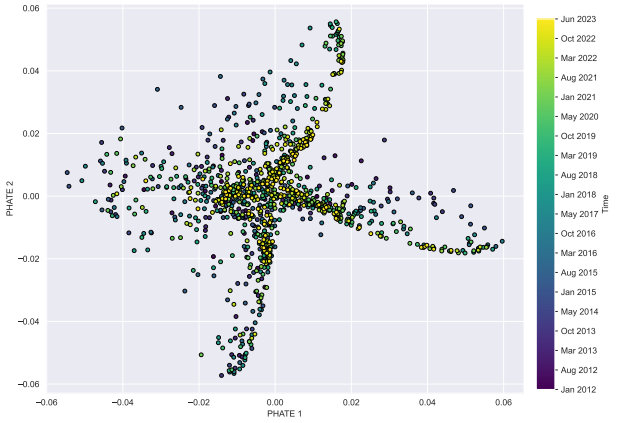
H. Case Study: arxiv-multi-4

"arxiv-multi-4" was an additional dataset chosen from arXiv for minimal category overlap, combining cryptography and security (cs.CR), fluid dynamics (physics.flu-dyn), differential geometry (math.DG), and probability (math.PR). These topics each represent a branch of the compass structure, clockwise starting from north. These embeddings make use of meso-scale adjustments using FAISS to rapidly compute and index an approximate 100-nearest neighbors graph, prior to applying PHATE. This dataset is of interest because of the separable tendril structure which visually shows that the *probability* topic (left branch of compass) is more interspersed with the other three topics. The diagram also shows how the multi-scale analysis reveals a topic split that correlates with the coloration by time chunks.



(a) The embedding reveals a global compass structure of the four arXiv category labels (clockwise from the "north" branch): *Cryptography/Security, Fluid Dynamics, Differential Geometry, and Probability*. Colors indicate meta-topic clusters.

(b) A lower cut height in the topic dendrogram gives a more fine-grained interpretation. There is a particularly interesting separation in the fluid dynamics topic which matches the temporal separation below.



(c) When the PHATE-Hellinger scatterplot is colored by time, some branches show temporal separability.

Figure 13. Hellinger-PHATE embeddings of arxiv-multi-4. Each point represents a single chunk topic. (a) The global structure shows category labels arranged by compass directions. (b) Finer granularity is revealed with a lower dendrogram cut height. (c) Temporal separability is evident in specific branches.

I. Document Interdisciplinarity Correlation with Topic Category Meta-Data

All datasets analyzed contained meta-data indicating topic categories. These topic categories were selected either by each document's authors, or manually post-processed and categorized. With a finite set of possible topic categories to choose from, it is not necessarily expected that a probabilistic topic model like our ensemble of LDA models would capture similar topics as the manually-labeled categories. However, as a sanity check, interdisciplinarity scores should be correlated with the number of categories that each respective author team has participated in, which is shown in Figure 14

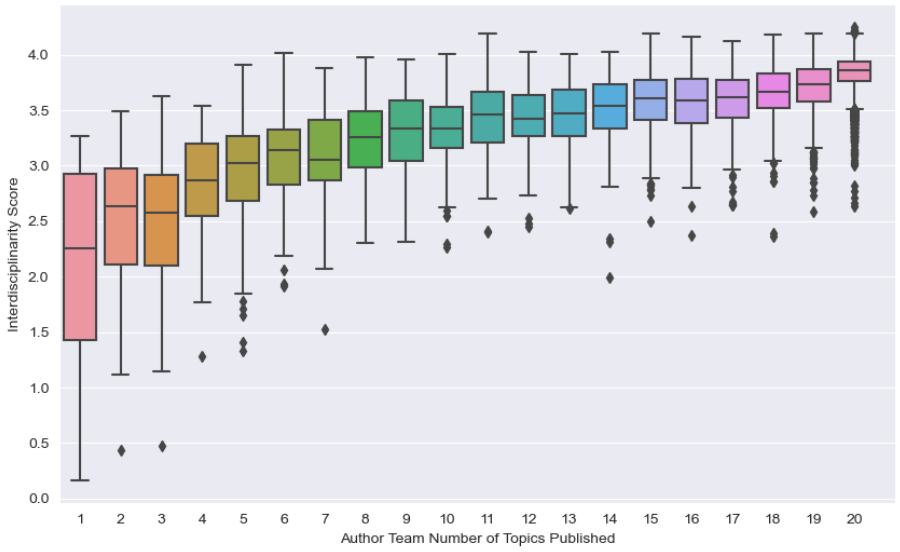
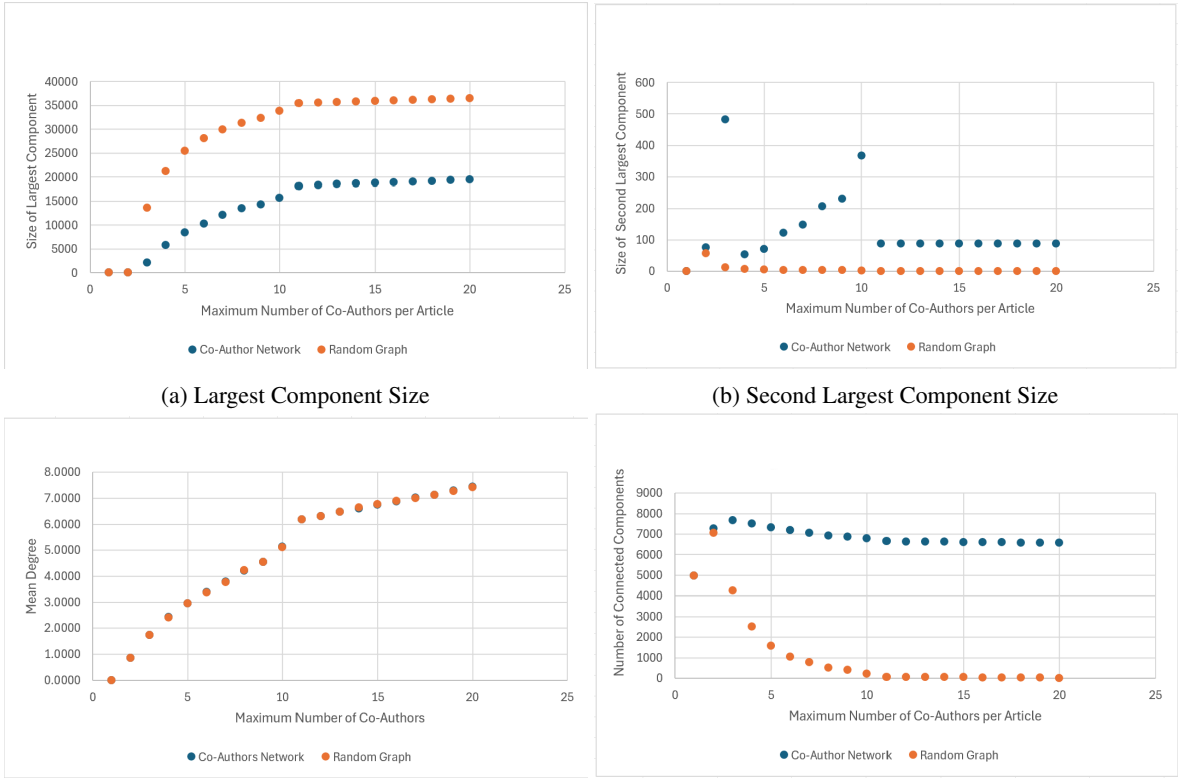


Figure 14. Interdisciplinarity box plot plotted against the number of category labels associated with that author team (cut height $h = 0.5$). This figure is one example, produced using 20 newsgroups data from the publicly-available library in python.

J. Network Percolation Experiment

J.1. Growth of Largest Connected Component



(a) Largest Component Size

(b) Second Largest Component Size

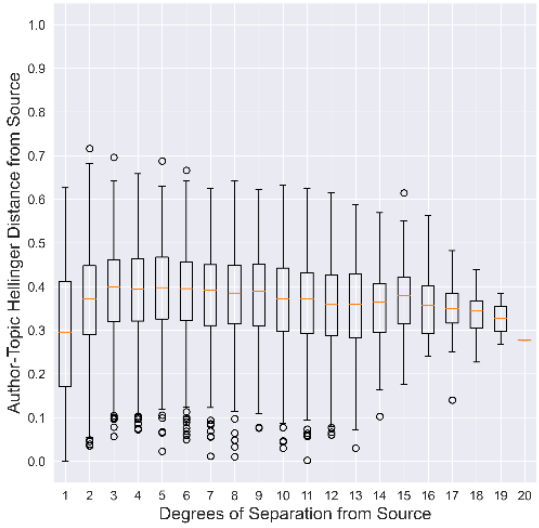
(c) Mean Degree

(d) Number of Components

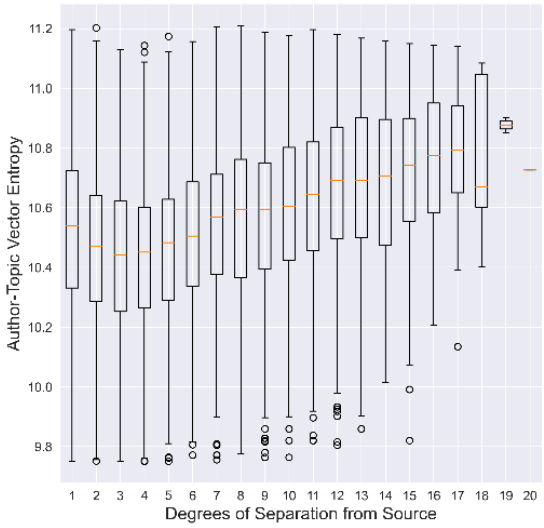
Figure 15. The mean degree of the OSTI co-author network is greater than 1. Networks with an average degree greater than 1 exhibit percolation behavior and contain a massive connected component that dominates the network. Restricting the document set based on a maximum number of co-authors affects the phase transition behavior of the large component size. The large component size of the OSTI co-author network does not grow as rapidly as in a random network of equivalent mean degree.

J.2. Small World Effect and Inter-Author Hellinger Distances

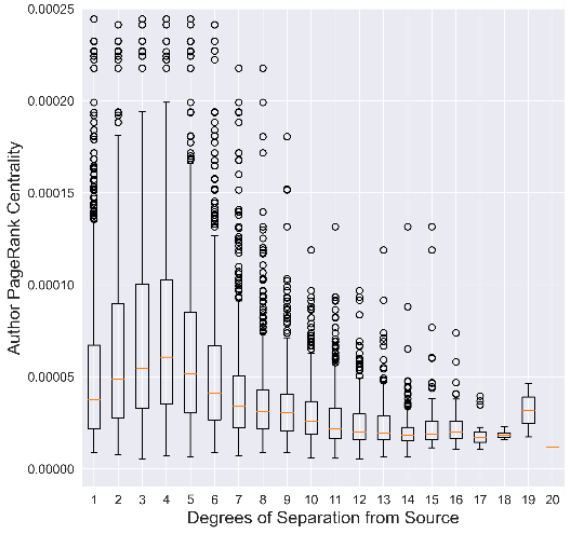
Figure 16 below shows a simple experiment which is meant to show how random walks through the co-author network suffer from the overwhelming central force of a massive connected component. The small world effect impacts the distances between authors in the topic space. Starting from a source author, paths were sampled, where at each step, t, it was required for the currently sampled author to be t steps away from the source author. Various metrics were computed. The topic space Hellinger distance between the source author and each sampled author interestingly increases for only a few steps, then decreases again. We see the reverse trend with author-topic distribution entropy. Central authors in the network, which are on average 3 to 4 steps away from a random source author, tend to be more concentrated in a particular topic. We know these sampled authors are central due to the PageRank centrality plot. The sampled path volumes show that paths up to 11 or 12 steps are reliable.



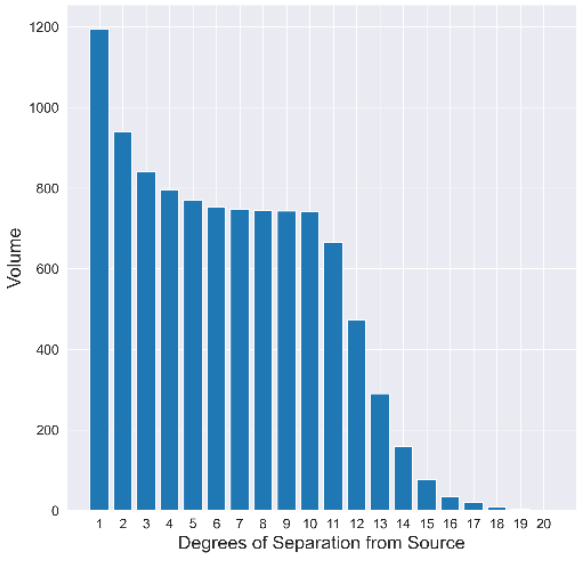
(a) Topic distance vs degree of separation.



(b) Author topic vector entropies.



(c) Author PageRank centrality.



(d) Sampled path volumes.

Figure 16. (a) Topic distance vs degree of separation, (b) Author topic vector entropies, (c) Author PageRank centrality, and (d) Sampled path volumes.