Community Detection with Hierarchical Topic Modelling

Mingyuan Zhang

National University of Singapore

Supervised by: Professor Wanjie Wang

22 April 2025

Abstract

Community detection and topic modeling are two fundamental techniques for uncovering structure in networked and textual data. In this paper, we propose an integrated framework that combines probabilistic community detection with hierarchical topic modeling. By applying community detection on a citation network followed by hierarchical Latent Dirichlet Allocation (hLDA) within each community, we reveal both structural and semantic patterns in the dataset. Our results show that incorporating textual node attributes significantly improves the interpretability of detected communities.

1 Introduction

Community detection and topic modeling are two widely used techniques for uncovering latent structure in complex datasets. Community detection aims to identify groups of nodes within a network that are more densely connected to each other than to nodes outside the group. This approach has been extensively applied in graph-based domains such as social networks, citation graphs, and biological systems to reveal functional clusters or communities. In contrast, topic modeling is a class of unsupervised learning methods used to uncover latent semantic themes from collections of text documents. By capturing word co-occurrence patterns, topic models provide compact and interpretable summaries of large corpora, revealing how terms cluster into topics and how topics are distributed across documents.

Although both community detection and topic modeling offer valuable insights—one into network structure, the other into textual semantics—they are typically applied in isolation. However,

many real-world networks contain both structural and textual information. For instance, in citation or co-authorship networks, each node represents a research paper and is associated with textual content such as a title or abstract. Traditional community detection algorithms operate solely on the network topology and disregard such node-level attributes, potentially resulting in communities that are structurally sound but semantically uninformative.

This observation motivates the integration of topic modeling with community detection. By jointly considering network structure and node-level text, we can enrich community analysis with semantic interpretation. Specifically, we propose a two-stage approach: (i) apply a probabilistic community detection algorithm to identify structural clusters of nodes, and (ii) perform hierarchical topic modeling on the textual content of each community to extract multi-level latent themes. The hierarchical structure enables finer-grained characterization of each community's topical landscape, allowing for comparative analysis across communities.

Formally, we consider a graph $G = (V, E)$, where V is a set of nodes and E is a set of edges. The goal of community detection is to partition V into subsets such that intra-community connections are dense and inter-community connections are sparse. In our setting, each node is also associated with a document, providing covariates that can be exploited during post hoc interpretation. Our research aims to address the following question: *How can textual node attributes be leveraged to enhance the interpretability of communities discovered via structural analysis?*

The contributions of this work are threefold. First, we conduct community detection on a citation network to identify structurally coherent groups of documents. Second, we apply a hierarchical topic model—specifically, hierarchical Latent Dirichlet Allocation (hLDA)—to the text within each community to uncover multi-level topic hierarchies. Third, we synthesize the results to generate thematic profiles for each community, offering an interpretable summary of both the network's structural and semantic organization.

2 Literature Review

2.1 Evolution of Topic Modeling

Early approaches to topic modeling trace back to techniques like Latent Semantic Analysis (LSA) in the 1990s. LSA utilized singular value decomposition to reduce the dimensionality of word-document co-occurrence matrices, revealing latent concepts that capture correlated terms (Deerwester et al. 1990). However, LSA lacked a probabilistic foundation, limiting its interpretability.

A significant advancement came with Latent Dirichlet Allocation (LDA), a generative probabilistic model representing each document as a mixture of topics, with each topic being a distribution over words (Blei, Ng, and Jordan 2003). LDA introduced a full generative model for

documents, incorporating Dirichlet priors for topic distributions, and addressed limitations of earlier models like probabilistic LSA.

Subsequent research focused on inference techniques and model extensions. Collapsed Gibbs sampling emerged as an effective inference method for LDA, allowing for practical estimation on large corpora and quantification of uncertainties (Griffiths and Steyvers 2004). To overcome the need to pre-specify the number of topics, non-parametric models like the Hierarchical Dirichlet Process (HDP) were introduced, enabling an unbounded number of topics (Teh et al. 2006).

Building upon these foundations, Hierarchical LDA (hLDA) was developed, utilizing the nested Chinese Restaurant Process as a prior to organize topics into a tree structure (Blei, Griffiths, and Jordan 2010). hLDA uncovers multi-level topic structures, mirroring the hierarchical organization of human knowledge, and provides both general and specific topics along with their relationships.

In recent years, neural topic models have further advanced the field. Neural approaches employ deep learning techniques, such as variational autoencoders, to learn topic distributions. For instance, Autoencoding Variational Inference for Topic Models (AVITM) leverages neural networks to infer topic vectors, achieving comparable accuracy to traditional methods with faster inference times (Srivastava and Sutton 2017). Despite these advances, hierarchical models like hLDA remain crucial for interpretability, as they explicitly organize topics in a human-understandable tree structure. In this paper, we utilize hLDA to capitalize on its ability to produce a hierarchy of topics, essential for interpreting network communities at multiple levels of granularity.

2.2 Community Detection in Networks

Community detection is a fundamental task in network science, aiming to identify groups of nodes—communities—that are more densely connected internally than with the rest of the network. Early work by Girvan and Newman introduced an algorithm based on iteratively removing edges with high betweenness centrality, effectively revealing community structures in complex systems (Girvan and M. E. Newman 2002).

Broadly, community detection methods can be categorized into: (i) Modularity-based methods, (ii) Spectral clustering, and (iii) Probabilistic models.

Modularity-based methods optimize a quality function called *modularity*, which measures the density of links inside communities compared to links between communities (M. E. Newman 2004). The Louvain algorithm is a notable example, employing a greedy optimization approach to maximize modularity efficiently in large networks (Blondel et al. 2008). However, modularity optimization has limitations, such as the resolution limit, which may prevent the detection of small communities within large networks (Fortunato and Barthélemy 2007).

Spectral clustering methods utilize the eigenvectors of matrices like the graph Laplacian or

modularity matrix to partition nodes. Newman proposed a spectral method using the leading eigenvector of the modularity matrix to detect community structures (M. E. Newman 2006). These methods are grounded in linear algebra and can provide insights into the network's structure, although they may require the number of communities to be specified in advance.

Probabilistic models, particularly the Stochastic Block Model (SBM), approach community detection as an inference problem on generative models of graphs. The SBM assumes that each node belongs to a latent group, and the probability of an edge between two nodes depends solely on their group memberships (Holland, Laskey, and Leinhardt 1983). Extensions of the SBM, such as those by Snijders and Nowicki (Snijders and Nowicki 1997) and Wasserman and Anderson (Wasserman and Anderson 1987), have introduced Bayesian frameworks and allowed for overlapping and hierarchical community structures.

Modern inference techniques for SBMs include Markov Chain Monte Carlo and variational methods. Tools like the `graph-tool` library implement efficient algorithms for fitting SBMs to large networks using Bayesian approaches (Tiago P Peixoto 2014a). In our context, applying SBM to a citation or co-authorship network can uncover groups of papers or authors that are densely connected, potentially corresponding to research communities or thematic clusters.

2.2.1 Stochastic Block Model

The Stochastic Block Model (SBM) is a widely used generative model for community detection in networks. It assumes that nodes are partitioned into communities, with the probability of an edge between two nodes depending solely on their community memberships. While SBM has been instrumental in understanding network structures, it has several limitations.

One notable issue is its assumption that all nodes within a community have similar connection probabilities, which does not account for degree heterogeneity. Real-world networks often exhibit broad degree distributions, with some nodes acting as hubs. The traditional SBM may incorrectly group high-degree nodes together, not because they belong to the same community, but due to their high connectivity. To address this, Karrer and Newman introduced the degree-corrected SBM, incorporating a degree parameter for each node. This extension allows the model to separate the effects of a node's degree from its community membership, improving community detection in networks with hubs (Karrer and M. E. J. Newman 2011).

Another limitation is that the SBM relies solely on network structure and does not incorporate node attributes. In networks where nodes have rich side information, such as textual content or metadata, purely structural communities might overlook important similarities between nodes. For instance, two papers from different communities might be topically similar, but an SBM would not group them together in the absence of a direct connection. This limitation motivates the integration of SBMs with content-based models.

These limitations highlight the need to integrate SBM with topic modeling. While SBM excels at detecting communities based on link patterns, topic modeling can provide an independent validation or interpretation using node attributes. In cases where SBM groups nodes purely based on structure, topic modeling can reveal if they are about different topics, indicating a potentially spurious community merge. Conversely, if SBM splits communities that are topically similar, topic modeling can uncover a higher-level relationship between those communities.

Given the aforementioned limitations, there is strong motivation to integrate network community detection with text mining. Community detection clusters documents based on connectivity, such as shared authors or citations, but without textual analysis, the thematic reasons for these groupings remain unclear. Applying topic modeling to the documents allows for the interpretation of each community in terms of prevalent themes.

This integration offers multiple benefits:

1. *Interpretability*: Communities can be labeled with informative topic keywords, aiding in understanding the thematic composition of each group.

2. *Validation*: A coherent topic profile within a community increases confidence in its meaningfulness. Conversely, a community with diverse topics may indicate structural grouping without thematic coherence.

3. *Discovering Inter-Community Relationships*: Analyzing topics across communities can reveal shared themes, suggesting higher-level groupings or potential future mergers.

In our approach, after obtaining communities and topics, we perform a semantic comparison of communities by constructing a topic-based similarity measure between them.

Recent literature has explored combining graph and content information for community detection, often referred to as ”attributed community detection” or ”content-aware clustering”. For example, Yang et al. developed CESNA, a model that incorporates node attributes into community assignment probabilities (Yang, McAuley, and Leskovec 2013). Our approach, however, maintains separate steps for community detection and topic modeling, integrating the results post hoc. This strategy offers simplicity and flexibility, allowing the use of specialized algorithms for each task and facilitating the combination of insights without developing a complex unified model.

3 Methodology

3.1 Community Detection Using the Stochastic Block Model

The initial phase of our methodology involves detecting communities within the document network by employing the Stochastic Block Model (SBM). We utilize the Bayesian inference framework

implemented in the `graph-tool` Python library (Tiago P. Peixoto 2014b), which efficiently infers the partition of nodes that best explains the observed network structure.

In this approach, the network is represented as a graph $G = (V, E)$, where V denotes the set of documents and E represents the connections between them (e.g., citations or co-authorships). To ensure meaningful community detection, we focus on the largest connected component of the graph, thereby excluding isolated nodes or trivial components that could skew the results.

The SBM inference process aims to maximize the posterior probability $P(z \mid G)$, where z denotes the community assignments for all nodes. This Bayesian framework allows for the automatic determination of an appropriate number of communities without the need for manual specification.

Post-inference, we observe that some communities identified are relatively small, which may not provide substantial insights for our analysis. To address this, we apply a filtering criterion, excluding communities comprising fewer than 30 documents. This threshold is chosen to focus on significant clusters that are more likely to represent coherent research themes.

It's important to note that at this stage, the community detection is based solely on the structural aspects of the network, without incorporating any textual content from the documents. Consequently, the identified communities are "structure-based", and their thematic coherence remains to be evaluated. The subsequent step in our methodology involves applying topic modeling techniques to these communities to uncover and interpret their underlying thematic structures.

3.2 Hierarchical LDA and the Nested Chinese Restaurant Process

Hierarchical Latent Dirichlet Allocation (hLDA) is an extension of LDA that models topics hierarchically rather than assuming a flat structure. Unlike LDA, which represents documents as mixtures of topics from a predefined set, hLDA allows topics to be arranged in a tree structure, where more general topics are found at higher levels, and more specific subtopics emerge deeper in the hierarchy.

The challenge in hierarchical topic modeling is determining the optimal tree structure for a given dataset. hLDA addresses this by using the nested Chinese Restaurant Process (nCRP) as a prior over tree structures, enabling a nonparametric approach that automatically discovers the number of topics and their hierarchical organization. Algorithm 1 describes the generative procss of hLDA.

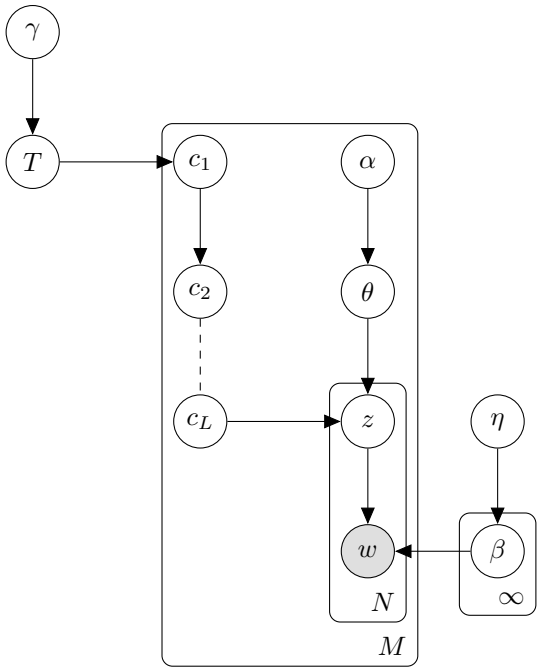

Figure 1: Graphical model of hLDA showing the hierarchical topic path generation via the nCRP and word generation through level-specific topics.

Algorithm 1 Generative Process for hLDA

Let c_1 be the **root restaurant**.
for each level $\ell \in \{2, \ldots, L\}$ **do**
    Draw a table from restaurant $c_{\ell-1}$ using CRP formula.
    Set c_ℓ to be the restaurant referred to by that table.
end for
Draw an L-dimensional topic proportion vector $\theta \sim \text{Dir}(\alpha)$.
for each word $n \in \{1, \ldots, N\}$ **do**
    Draw topic assignment $z_n \sim \text{Mult}(\theta)$.
    Draw word w_n from the topic associated with restaurant c_{z_n}.
end for

3.2.1 Priors

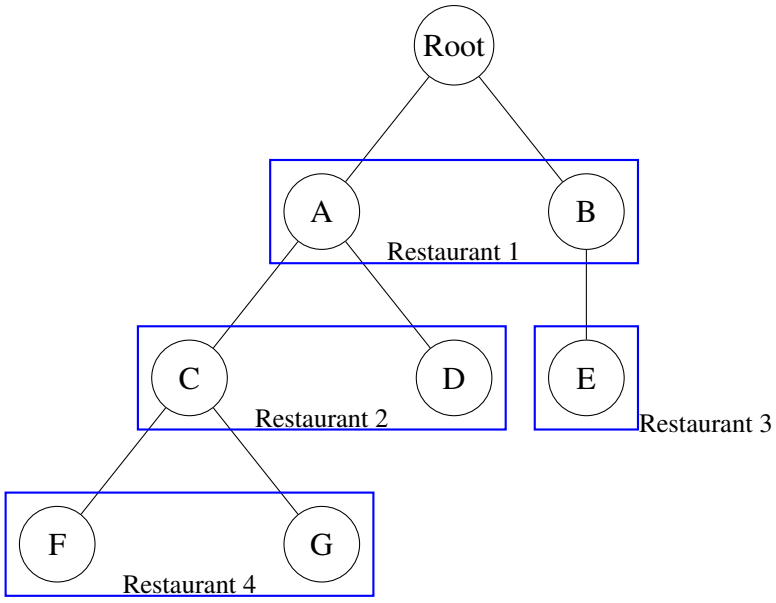

Figure 2: Illustration of Nested Chinese Restaurant Process (nCRP)

Chinese Restaurant Process(CRP) consider customers entering an infinite-table Chinese restaurant. Customer n sits at an *occupied* table i or opens a *new* table according to

$$\begin{aligned} P(\text{table } i) &= \frac{n_i}{\gamma + n - 1}, \\ P(\text{new table}) &= \frac{\gamma}{\gamma + n - 1} \end{aligned} \tag{1}$$

where n_i is the number already at i and γ is a concentration parameter. Equation(1) induces a power law cluster size distribution while leaving the cluster count unbounded.

Nested CRP (nCRP) stacks CRPs recursively. Each table at level ℓ points to a *child restaurant* at level $\ell+1$. A document (customer) starts at the root restaurant, chooses a table via (1), moves to the linked child restaurant, and repeats for L levels. The result is a random path $c_d = (t_{d,1}, \ldots, t_{d,L})$ and a random subtree shared across documents (Figure 2).

3.2.2 Gibbs Sampling inference

Following (Griffiths and Steyvers 2004), Inference in hLDA is performed using Gibbs sampling, which iteratively updates:

- Topic assignment z for each word in a document.

- Path assignment c for each document.

Each update step integrates information from the entire corpus, refining the hierarchical structure over time.

3.2.3 Sampling z (Level Assignment)

The probability of assigning a word w_n to a level z_n in document m follows:

$$P(z_n = \ell \mid \mathbf{z}_{-n}, \mathbf{w}) \propto \frac{n_{c_m,\ell}^{(w_n)} + \eta}{n_{c_m,\ell}^{(\cdot)} + W\eta} \times \frac{n_{c_m,\ell}^{(m)} + \alpha}{n_{c_m,\cdot}^{(m)} + L\alpha} \tag{2}$$

- $\mathbf{z}_{-n}$: Assignments of all z_k except for z_n (i.e., $k \neq n$).
- $n_{c_m,\ell}^{(w_n)}$: Count of words assigned to level ℓ that match w_n, excluding w_n.
- $n_{c_m,\ell}^{(\cdot)}$: Total number of words assigned to level ℓ, excluding w_n.
- $n_{c_m,\ell}^{(m)}$: Count of words from document m assigned to level ℓ, excluding w_n.
- $n_{c_m,\cdot}^{(m)}$: Total words in document m, excluding w_n.
- α, η: Smoothing parameters.
- L: Maximum number of levels

3.2.4 Sampling c (Path Assignment)

For a document d, we sample a new path c_d via:

$$p(c_m \mid \mathbf{w}, \mathbf{c}_{-m}, Z) \propto p(\mathbf{w}_m \mid \mathbf{c}, \mathbf{w}_{-m}, Z) p(c_m \mid \mathbf{c}_{-m}), \tag{3}$$

The document-likelihood term is:

$$p(\mathbf{w}_m \mid \mathbf{c}, \mathbf{w}_{-m}, \mathbf{z}) = \prod_{\ell=1}^{L} \left(\frac{\Gamma(n_{c_m,\ell,-m}^{(\cdot)} + W\eta)}{\prod_w \Gamma(n_{c_m,\ell,-m}^{(w)} + \eta)} \frac{\prod_w \Gamma(n_{c_m,\ell,-m}^{(w)} + n_{c_m,\ell,m}^{(w)} + \eta)}{\Gamma(n_{c_m,\ell,-m}^{(\cdot)} + n_{c_m,\ell,m}^{(\cdot)} + W\eta)} \right), \tag{4}$$

- $\mathbf{w}_m$: Words in document m.
- $\mathbf{w}_{-m}$: Words in all documents except m.
- $\mathbf{c}_{-m}$: Path assignments for all documents except m.
- $n_{c_m,\ell,-m}^{(w)}$: Count of word w assigned to topic ℓ, excluding document m.

- $n^{(w)}_{c_m,\ell,m}$: Count of word w assigned to topic ℓ within document m.
- $n^{(\cdot)}_{c_m,\ell,-m}$: Total number of words assigned to topic ℓ, excluding document m.
- $n^{(\cdot)}_{c_m,\ell,m}$: Total number of words assigned to topic ℓ within document m.

Using the property of Γ-functions:

$$\frac{\Gamma(x+n)}{\Gamma(x)} \;=\; \prod_{i=0}^{n-1}(x+i).$$

The document-likelihood term can be simplified to:

$$p(\mathbf{w}_m \mid \mathbf{c}, \mathbf{w}_{-m}, \mathbf{z}) \;=\; \prod_{\ell=1}^{L}\prod_{w=1}^{W}\prod_{i=0}^{n^{(m)}_{c_m,\ell,w}-1}\frac{\eta + n^{(-w)}_{c_m,\ell,-m} + i}{\eta W + n^{(\cdot)}_{c_m,\ell,-m} + i}. \tag{5}$$

In words, the process follows the following algorithm.

Algorithm 2 Document-Likelihood Computation

```
log_prob ← 0
for each level ℓ in the topic tree do
    for each vocabulary word w ∈ {1, ..., W} do
        for each occurrence i ∈ {0, ..., n^(m)_{c_m,ℓ,w} − 1} in document m do
            log_prob += log( (η + n^(−w)_{c_m,ℓ,−m} + i) / (ηW + n^(·)_{c_m,ℓ,−m} + i) )
        end for
    end for
end for
```

3.2.5 Hyperparameters

The result of hLDA is significantly influenced by three key hyperparameters:

- α (Document-topic distribution parameter)
- η (Topic-word distribution parameter)
- γ (nCRP parameter)

The hyperparameter α is the Dirichlet prior on the document-topic distribution θ_d. It determines how spread out the topic proportions are in each document. Fη controls the Dirichlet prior on the topic-word distribution β, affecting how words are assigned to topics. Finally, the hyperparameter

γ comes from the nCRP and determines how likely a document is to choose a new branch in the topic hierarchy.

Hyperparameter	Low Value	High Value
α (document-topic)	Few dominant topics per doc	Many topics per doc
η (topic-word)	Distinct, sharp topics	Overlapping, mixed topics
γ (tree prior)	Fewer topic branches	Many topic branches

Table 1: Effects of Hyperparameter Tuning in hLDA

Choosing appropriate values for α, η, and γ is crucial for balancing topic diversity, specificity, and hierarchical depth in hLDA.

Now, having both community assignments (from SBM) and topic assignments (from hLDA), the final methodological step is to integrate these results to derive insights about communities using topics.

3.3 Integration of SBM and hLDA Results

The integration process comprises two main components: (1) analyzing the internal thematic structure of each community, and (2) comparing different communities based on their topic distributions.

3.3.1 Community-Level Topic Aggregation

For each community C identified by the SBM, we aggregate the hierarchical topic distributions from hLDA as follows:

1. Initialize a vector v_C of length equal to the total number of topics across all levels, with entries set to zero.

2. For each document $d \in C$, retrieve its topic distribution $[z_{d,1}, z_{d,2}, \ldots, z_{d,L}]$. For each level ℓ, increment $v_C[z_{d,\ell}]$ by the corresponding topic proportion.

3. Normalize v_C to obtain a probability distribution $\hat{v}_C$ over topics.

4. Discard topics in $\hat{v}_C$ with negligible probabilities (e.g., below 0.01%) to focus on significant themes.

5. Map the remaining topics back into the hierarchical tree structure to derive a pruned subtree representing the community's main themes.

This aggregation effectively summarizes the thematic composition of each community. For instance, if Community A predominantly contains documents on quantum physics, the aggregated topic distribution might highlight a path such as Science $\rightarrow$ Physics $\rightarrow$ Quantum Physics, with corresponding proportions indicating the community's focus at each hierarchical level.

3.3.2 Inter-Community Comparison via Topic Clusters

To compare communities based on their thematic profiles, we construct a topic similarity network:

1. Represent leaf topic discovered by hLDA as a node in the network.
2. Connect two topic nodes with an edge if their word distributions have a cosine similarity above a predefined threshold, indicating semantic relatedness.
3. Apply the SBM to this topic network to identify clusters of related topics, resulting in a set of broader thematic groups.

We then:

1. Initialize a vector T_C of length equal to the total number of topic clusters found by SBM in the topic-topic network.
2. For each document $d \in C$, retrieve the leaf topic's topic cluster, t_d, in the topic-topic network. increment $T_C[t_d]$ by the corresponding topic proportion.
3. Normalize these sums to create a community topic distribution vector.

By comparing these topic distribution vectors across communities, we can assess thematic similarities and differences. Communities with similar profiles may cover related subject areas, even if they are structurally distinct in the citation network. Conversely, distinct profiles suggest divergent research focuses. In our approach, we utilize only the leaf nodes from the hLDA-generated topic hierarchy. We interpret the hLDA process as a filtering mechanism, where general topics at higher levels are excluded, and the most specific topics, located at the leaves, are retained. These leaf-level topics best represent each document's content, providing a more precise thematic characterization. This strategy also reduces the number of topics to consider, simplifying subsequent analyses.

This methodology enables a nuanced understanding of the thematic landscape of the document network, facilitating the identification of both cohesive research areas and potential interdisciplinary connections.

4 Data Description

The dataset for this study is derived from the Annals of Applied Statistics (AoAS) publication records. Each node in the network is a document, and edges represent two types of relationships:

- **Direct citation**: If paper A cites paper B (within the AoAS corpus), we create an edge between A and B.
- **Co-authorship linkage**: If paper A and paper B share one or more authors, we create an edge between A and B. This links documents that are related via authorship (even if they do not cite one another).

We treat the edges as undirected for community detection (the network is made undirected by considering a citation or co-authorship as an undirected link). The rationale for including both citations and co-authorship is to capture a broader notion of "relatedness" between papers: citation links indicate topical relevance or influence, while co-authorship links indicate research collaboration or common scholarly domain.

From the AoAS records, we extracted a total of 3248 documents (nodes). The total number of unique authors across these papers is 3607. We note that many authors have written multiple papers, creating a dense co-authorship network among their papers.

Using the above criteria, we constructed the edge list. An edge was added for each direct citation within this set and for each pair of papers with a common author. This yielded a network with $|V| = 3248$ nodes and $|E| = 21356$ edges. On average, each paper is connected to about 13 others, though the degree distribution is skewed: some papers are connected to dozens of others, while some have only one or two connections.

We found that the graph is not fully connected. There is one giant connected component containing the majority of the papers, and a number of smaller components. The largest connected component has most of the nodes. We focus our community detection on the largest component. Smaller isolated components were ignored for community analysis since they typically represent trivial cases.

For each document node, we also collected its text for topic modeling. We used the abstract as the document text. We performed standard text preprocessing: lowercasing, removing stop words, and stemming. The vocabulary was then built. In total, there are more than 10000 unique vocabularies.

5 Results

5.1 Community Detection Results (SBM)

We applied Stochastic Block Model (SBM) inference to the largest connected component of the document-document network, which included 3215 out of 3248 papers. The SBM produced numerous communities, many of which were quite small. To focus on interpretable and statistically robust clusters, we filtered out communities with fewer than 30 papers. After this thresholding step, we retained *43 communities* for further analysis.

These 43 communities varied in size, with the smallest containing 31 documents (as determined by our filter) and the largest containing 112. The average community size was 56 papers. Figure 3 shows the distribution of the sizes of the communities detected.

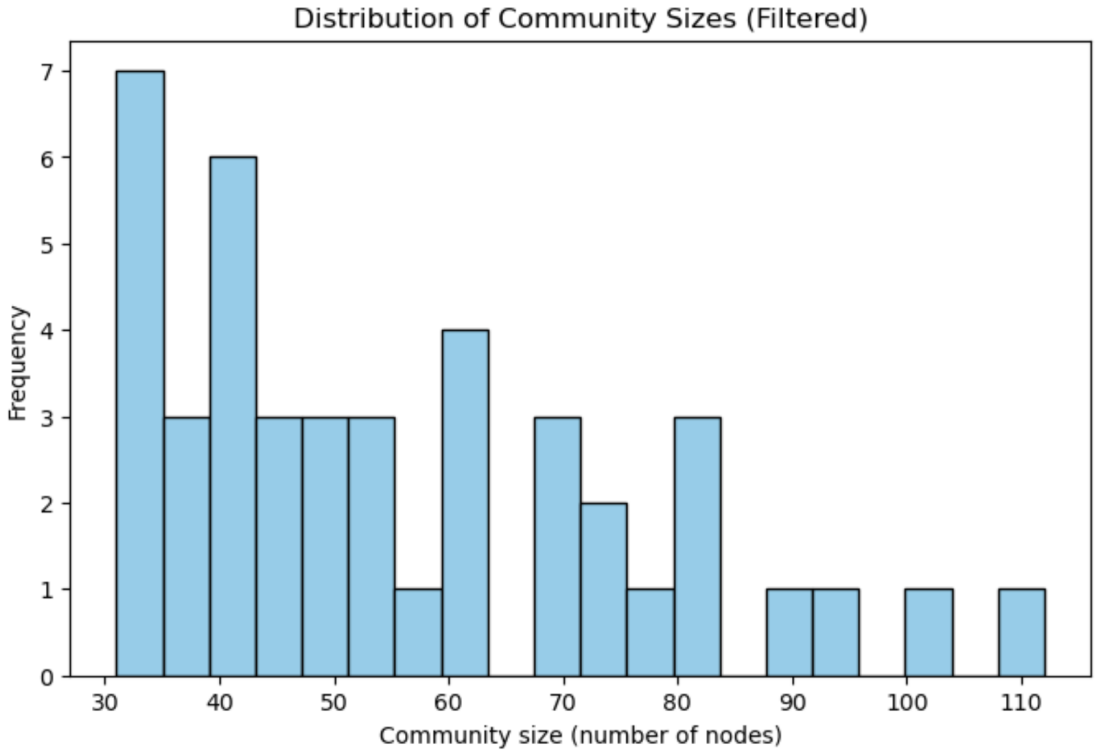

Figure 3: Community distribution

The existence of a community with over 100 papers reflects a densely interconnected subfield, while the presence of many communities of size 30–50 suggests more focused or niche research areas.

Importantly, these communities were detected solely from the network structure—based on citation and co-authorship information—with no access to document content.

5.2 Topic Modeling Results (hLDA)

We applied hierarchical Latent Dirichlet Allocation (hLDA) to the filtered corpus of 3215 documents using abstracts. The model was configured to allow a maximum depth of $L = 4$, and the concentration parameter γ for the nested Chinese Restaurant Process (nCRP) was tuned to balance depth and breadth of the inferred topic tree. After convergence of the collapsed Gibbs sampler, the model yielded a total of 669 distinct topics across all levels.

The distribution of topics by level was as follows:

- Level 1: 1 topic
- Level 2: 28 topics
- Level 3: 113 topics
- Level 4: 527 topics

Level 1 contained exactly one root topic, representing a catch-all node for the corpus—interpreted broadly as "All Applied Statistics." At level 2, the model discovered 28 broad thematic areas that reflect major subfields within applied statistics. Examination of the top words for these level-2 topics revealed recognizable domains such as bioinformatics, financial statistics, survey methodology, Bayesian inference, and machine learning. Level 3 refined these broad categories into 113 more granular topics. For example, the level-2 bioinformatics topic was subdivided into child topics including gene expression analysis, genome-wide association studies, and protein interaction networks. Level 4, the most specific layer, accounted for 527 topics. These topics often focused on specialized methods or narrowly scoped applications.

The hierarchical structure inferred by hLDA reveals semantically coherent clusters across multiple resolutions. A representative portion of the tree can be found in the appendix. This hierarchy illustrates that hLDA successfully discovered meaningful topic clusters at multiple levels of granularity, many of which align well with established subfields in applied statistics.

In conclusion, the hLDA yielded a comprehensive topic structure. We will now describe how these topics manifest within the previously found communities and how communities differ in their topic make-up.

5.3 Integrated Analysis: Community Thematic Profile for Community 173

To illustrate how structural and semantic analysis can be combined, we focus on Community 173—the largest community identified by the Stochastic Block Model. This community contains 112 documents and demonstrates a rich and diverse thematic composition. Using the procedure

described in Section 3.3, we aggregated the topic distributions of all documents in Community 173 and extracted its dominant topic path structure.

Figure 4 displays a truncated version of the hierarchical topic tree for Community 173. For readability, only the topics with the highest probabilities at each level are shown. This selective visualization is necessary due to the size of the full topic tree.

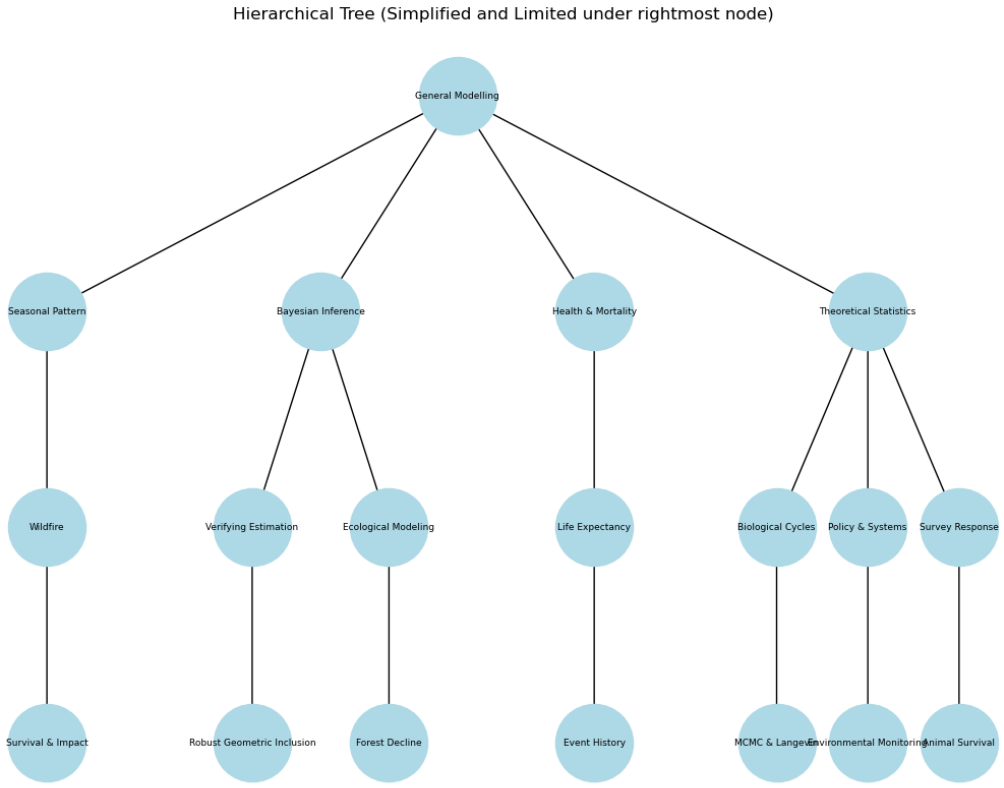

Figure 4: Hierarchical Tree of the largest Community, 172

From the visualization, we observe that Community 173 is anchored under a general "Modelling" root topic, which branches into four high-level domains: Seasonal Pattern, Bayesian Inference, Health and Mortality, and Theoretical Statistics. Each of these top-level themes gives rise to specific subfields. For example:

- The "Seasonal Pattern" branch narrows down to topics on "Wildfire" and ultimately "Survival and Impact Analysis."
- The "Bayesian Inference" path includes subtopics on "Verifying Estimation Procedures" and "Ecological Modeling," each of which branches into fine-grained applications such as "Robust Geometric Inclusion" and "Forest Decline."

- The "Health and Mortality" theme transitions into topics on "Life Expectancy" and further to "Event History and Survival Data."
- The "Theoretical Statistics" branch is particularly diverse, encompassing areas such as "Biological Cycles," "Policy and Systems," and "Survey Response," which themselves connect to specific technical themes like "MCMC and Langevin Sampling," "Environmental Monitoring," and "Animal Survival."

This structure highlights both the thematic richness and diversity within Community 173. While Bayesian computation and survival modeling emerge as central threads, the presence of multiple deep topic paths suggests that the community spans several interconnected subfields, including environmental statistics, theoretical modeling, and applied health analytics.

It is important to acknowledge the limitations of this visualization. Because hLDA produces a deep and wide tree with hundreds of leaf nodes, we only display the top-ranked topics in terms of word probability mass. While this approach enhances interpretability, it may omit lower-weight but still meaningful thematic branches. Future work could explore alternative visualization strategies such as interactive topic trees or dimensionality-reduced semantic maps to capture broader aspects of community semantics.

5.4 Semantic Comparison of Communities

In addition to assigning thematic labels to individual document communities, we sought to understand how similar these communities are to one another in terms of semantic content. Using the 472 topics from hLDA, we first constructed a topic-topic network based on cosine similarity between topic word distributions. We then applied Stochastic Block Modeling (SBM) to this network, identifying 21 higher-level *topic clusters*. Each document community was represented as a 21-dimensional probability vector over these topic clusters—effectively capturing the semantic fingerprint of each community.

We then computed pairwise cosine similarities between these community vectors, yielding a community-community similarity matrix that reflects topical proximity, rather than network linkage. To visualize this semantic landscape, we constructed a graph where nodes represent communities, and edges connect pairs with high topic similarity. Community detection on this semantic graph revealed higher-order groupings—clusters of communities with shared topical foundations.

Figure 5 presents a detailed comparison between two selected communities—Community 172 and Community 1047—through pie charts of their dominant topic distributions. Both communities are reasonably concentrated, but they differ significantly in thematic emphasis.

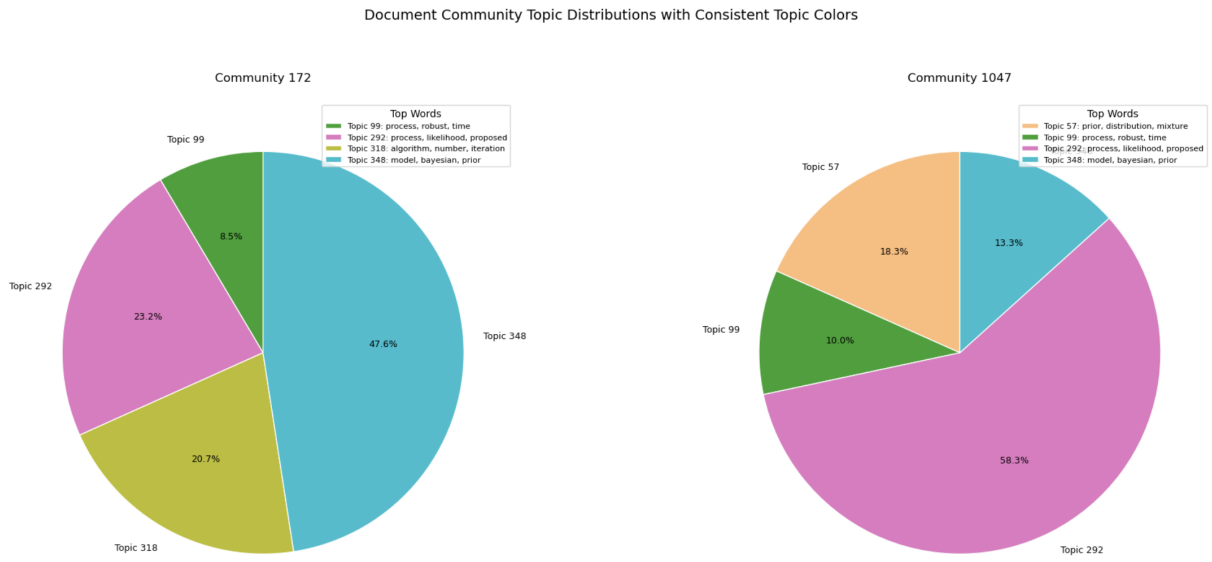

Figure 5: Topic composition of two document communities, using consistent topic color-coding. Each slice represents the proportion of words in the community assigned to a given topic.

- **Community 172** is heavily focused on Topic 348 (47.6%), which is defined by the keywords *model, bayesian, prior*. This indicates a strong Bayesian methodological core. The remaining distribution includes Topic 318 (20.7%; *algorithm, number, iteration*), Topic 292 (23.2%; *process, likelihood, proposed*), and a smaller contribution from Topic 99 (8.5%; *process, robust, time*). The presence of Topics 318 and 292 suggests a focus on algorithmic and likelihood-based inference methods, aligning with a computational-statistics orientation.

- **Community 1047**, on the other hand, is dominated by Topic 292 (58.3%), which overlaps with Community 172 but to a much greater extent. This suggests that Community 1047 shares some methodological interests with Community 172—especially in likelihood-based modeling—but with a different emphasis. It also includes Topic 57 (18.3%; *prior, distribution, mixture*), which is not present in Community 172, indicating a stronger focus on mixture models and probabilistic modeling. The presence of Topics 99 (10.0%) and 348 (13.3%) further reinforces the Bayesian influence, but to a lesser extent than in Community 172.

Despite having overlapping topics (notably 292, 348, and 99), these two communities differ in their dominant emphases: Community 172 leans more toward Bayesian computational methods and iterative algorithms, while Community 1047 places more weight on mixture modeling and likelihood-based inference.

Across all 43 communities, we observed that this topic-based representation revealed semantic relationships not visible from network structure alone. For example:

6 Discussion

6.1 Insights Gained from Integrating Text and Network Analysis

The integration of community detection and topic modeling has produced significantly richer insights than either method could provide in isolation. Community detection revealed the structural organization of the research network—how papers are grouped based on citation and co-authorship links—while topic modeling uncovered the semantic content of these documents. When combined, the two methods allow us to make detailed assertions such as: "AoAS contains a genomics-focused community (Community 5), a finance-oriented community (Community 12), and two distinct but topically similar communities in clinical statistics."

A major insight is the importance of incorporating textual information for interpreting communities. Without topic modeling, Community 5 would appear as a group of 95 inter-citing papers; identifying its focus would require manual inspection. With hLDA, we determined that its dominant theme is genomics, based on both topic keywords and hierarchical structure. More broadly, every community in the network was assigned a descriptive thematic label, transforming an abstract partitioning into a meaningful map of subfields in applied statistics. This directly addresses a common drawback of graph-based community detection: that the output is structurally valid but semantically ambiguous. Here, we use topic modeling to anchor network structure in topical content.

The hierarchical nature of hLDA further enhances interpretability. In many cases, the top topics associated with a community descended from a shared level-2 topic, indicating that the community is entirely situated within a broader research domain. In contrast, some communities drew significant weight from multiple high-level branches of the hierarchy, signaling interdisciplinary content. For example, we observed a community spanning both Bayesian inference and machine learning. Its topic distribution included significant contributions from a "Bayesian Methods" branch as well as a "Machine Learning" branch. This revealed the group's focus on Bayesian approaches to machine learning problems—an insight that would be obscured in flat topic models without hierarchical structure.

From the perspective of topic modeling, the community structure offers a form of external validation. Topics heavily concentrated within a single community (and absent in others) suggest that they represent specialized subfields well captured by SBM. In contrast, topics that are distributed across many communities point to common methodologies or general techniques. For instance,

topics related to “regression modeling” or “likelihood inference” appeared across a wide range of communities, reflecting their foundational role in applied statistics. In this way, topic distribution across communities acts as a lens for identifying both ubiquitous methods and tightly clustered thematic areas.

Together, network and semantic structures provide a multi-faceted view of the scientific landscape. Community detection organizes papers by collaboration and citation, while topic modeling reveals their intellectual content. Their integration allows not only for labeling communities with descriptive titles but also for identifying interdisciplinary bridges, validating subfield specialization, and contextualizing research themes within broader hierarchies. This synergy transforms disconnected outputs into a coherent narrative about the structure and content of scholarly discourse.

6.2 Challenges and Limitations

6.2.1 Stochastic Variability and Reproducibility

Both the Stochastic Block Model (SBM) and hierarchical Latent Dirichlet Allocation (hLDA) are generative probabilistic models whose outputs can vary across runs due to their reliance on sampling-based inference. In particular, hLDA with collapsed Gibbs sampling introduces variability in topic tree structures and level assignments. While general trends tend to remain stable, details such as the number of topics at lower levels or precise community boundaries may differ between runs. Selecting a representative result is non-trivial in the absence of ground truth, and evaluating multiple runs for consistency can be time-consuming. Furthermore, reliable metrics to assess the quality of unsupervised clusters or topic assignments are not always available, posing challenges for reproducibility.

6.2.2 Evaluation Without Ground Truth

Another challenge is the lack of ground truth labels for communities or topics. As a result, evaluation relied on qualitative methods such as examining document titles, subfield associations, and topic-word coherence. While internal consistency was strong—documents within communities shared similar topic distributions—external validation remains difficult. Some community-topic alignments matched known special issues or subdomains, but comprehensive verification would require expert input or curated metadata.

6.2.3 Modularity of the Integration Approach

Finally, the analysis follows a two-step pipeline: first detecting communities from network structure, then applying topic modeling for interpretation. This modular design allowed us to leverage specialized tools for each task, but it also introduces limitations. Because community detection used only structural information, some papers may be grouped together despite differing in content. Topic modeling helped interpret but not revise these assignments. A more integrated approach—such as a joint model where topics influence edge probabilities—could yield communities more aligned with content. While such models are more complex, they offer a promising direction for future refinement. An intermediate solution might involve post hoc adjustment: splitting or merging communities based on topic similarity to improve coherence.

7 Conclusion

This study presents an integrated approach to analyzing document networks by combining community detection through the Stochastic Block Model with hierarchical topic modeling via hLDA. Applied to a coauthorship-citation network from the *Annals of Applied Statistics*, this framework uncovered 43 structural communities and 669 topics arranged in a meaningful hierarchy. By mapping topic distributions to communities, we were able to assign clear semantic labels to each group, providing insight into the research landscape—spanning genomics, finance, environmental statistics, and methodology.

The integration offered several key benefits. It enhanced the interpretability of structural communities, uncovered hidden thematic relationships between disconnected groups, and revealed higher-order clusters through topic-based similarity. The hierarchical nature of the topics aligned well with the multi-scale organization of communities, supporting both broad and fine-grained analysis. Our findings demonstrate that structural proximity often correlates with topical coherence, and that combining structure with content yields insights not accessible through either in isolation.

Beyond the AoAS dataset, this framework is widely applicable to other domains involving networks with associated text—such as academic corpora, social networks, or collaboration graphs. It can support recommendation systems, thematic mapping of research fields, or editorial planning. Future work may explore temporal dynamics, joint models that couple structure and content during inference, or user-guided refinement of results. Despite its modular nature, the current two-step pipeline already demonstrates the value of integrating community detection and topic modeling to better understand complex, multi-modal networks.

8. References

Blei, David M, Thomas L Griffiths, and Michael I Jordan (2010). "Nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies". In: *Journal of the ACM (JACM)*. Vol. 57. 2. ACM New York, NY, USA, pp. 1–30.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3, pp. 993–1022.

Blondel, Vincent D et al. (2008). "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008.

Deerwester, Scott et al. (1990). "Indexing by latent semantic analysis". In: *Journal of the American Society for Information Science* 41.6, pp. 391–407.

Fortunato, Santo and Marc Barthélemy (2007). "Resolution limit in community detection". In: *Proceedings of the National Academy of Sciences* 104.1, pp. 36–41.

Girvan, Michelle and Mark EJ Newman (2002). "Community structure in social and biological networks". In: *Proceedings of the National Academy of Sciences* 99.12, pp. 7821–7826.

Griffiths, Thomas L and Mark Steyvers (2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.suppl 1, pp. 5228–5235.

Holland, Paul W, Kathryn Blackmond Laskey, and Samuel Leinhardt (1983). "Stochastic blockmodels: First steps". In: *Social Networks* 5.2, pp. 109–137.

Karrer, Brian and M E J Newman (2011). "Stochastic blockmodels and community structure in networks". In: *Physical Review E* 83.1, p. 016107.

Newman, Mark EJ (2004). "Fast algorithm for detecting community structure in networks". In: *Physical Review E* 69.6, p. 066133.

— (2006). "Modularity and community structure in networks". In: *Proceedings of the National Academy of Sciences* 103.23, pp. 8577–8582.

Peixoto, Tiago P (2014a). "Hierarchical block structures and high-resolution model selection in large networks". In: *Physical Review X* 4.1, p. 011047.

— (2014b). "The graph-tool python library". In: *figshare*. DOI: 10.6084/m9.figshare.1164194. URL: https://figshare.com/articles/dataset/graph_tool/1164194.

Snijders, Tom AB and Krzysztof Nowicki (1997). "Estimation and prediction for stochastic blockmodels for graphs with latent block structure". In: *Journal of Classification* 14.1, pp. 75–100.

Srivastava, Akash and Charles Sutton (2017). "Autoencoding variational inference for topic models". In: *5th International Conference on Learning Representations (ICLR)*.

Teh, Yee Whye et al. (2006). "Hierarchical Dirichlet processes". In: *Journal of the American Statistical Association* 101.476, pp. 1566–1581.

Wasserman, Stanley and Carolyn Anderson (1987). "Stochastic a posteriori blockmodels: Construction and assessment". In: *Social Networks* 9.1, pp. 1–36.

Yang, Jaewon, Julian McAuley, and Jure Leskovec (2013). "Community detection in networks with node attributes". In: *2013 IEEE 13th International Conference on Data Mining*. IEEE, pp. 1151–1156.

8 Appendix

8.1 Inferred Sub-Hiarerchical Tree

Topic 0: General Modelling

- **Topic 14: Seasonal Pattern**
  - Topic 366: Wildfire
    - Topic 636: Survival and Impact Analysis
- **Topic 13: Bayesian Inference with Markov Chains**
  - Topic 122: Verifying Estimation Procedures
    - Topic 346: Robust Geometric Inclusion Modeling
  - Topic 92: Ecological Modeling – Beetle and Forest Dynamics
    - Topic 651: Exponential Forest Decline Modeling
- **Topic 141: Health and Mortality Studies**
  - Topic 79: Life Expectancy and Disability Estimation
    - Topic 602: Event History and Survival Data
- **Topic 8: Theoretical Statistics and Estimation**
  - Topic 250: Biological Cycles and Circular Regression
    - Topic 643: MCMC and Langevin Sampling Algorithms
  - Topic 394: Policy and State-Dependent Systems
    - Topic 642: Environmental Monitoring and Ozone Trend
  - Topic 328: Survey Response Styles and Attitudes
    - Topic 689: Animal Survival and Ecological Capture

- Topic 396: Statistical Independence and Transformation
  - Topic 376: Particle Filtering and Dynamic State Estimation
- Topic 151: Brain Tissue Classification with Markov Fields
  - Topic 686: Bayesian Spatial Regression for Imaging
- Topic 126: Survey Estimation in Small Domains
  - Topic 80: Spatial Autoregressions and Variogram Analysis
- Topic 211: Saddlepoint Approximation and Density Modeling
  - Topic 515: Joint Distributions and Recursive Likelihood Computation