Exploring Trends and Venue Classification in the DBLP Citation Network

[Adam Johnson, 2031383] - [Bryan Pedroza, 2287321] - [Jason Quach, 2135612]

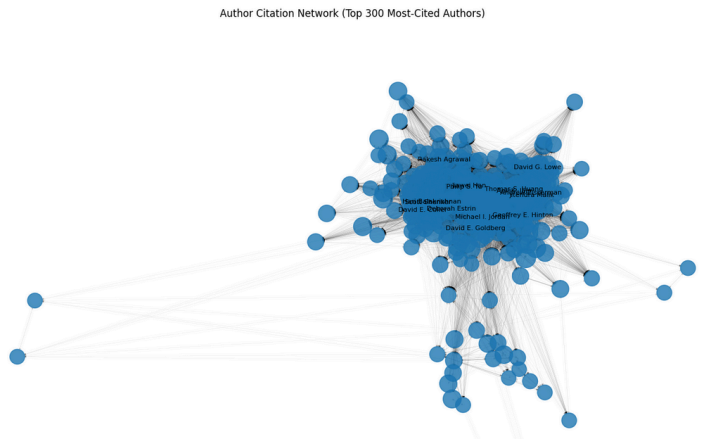
Introduction

For this project, we divided our work across three major components: Classification, Clustering, and Anomaly Detection. The data pipeline involved loading and cleaning the DBLP dataset, generating the TF-IDF features, and preparing the venue-level dataset for further analysis (as in the above paragraph). The exploratory analysis focused on identifying trends across venues. The classification component constructed balanced paper–venue datasets and developed models to compare how well different algorithms could predict venue labels, including a temporal version that trains on earlier years and tests on later ones. The clustering component applied K-means, hierarchical clustering, and DBSCAN to the PCA-scaled TF-IDF venue data, evaluating cluster quality with SSE, silhouette scores, and a penalized silhouette metric, and ultimately revealing stable, semantically coherent venue groups across algorithms. The anomaly detection component detects unusual patterns in papers, authors, and venue-years using distance-based methods.

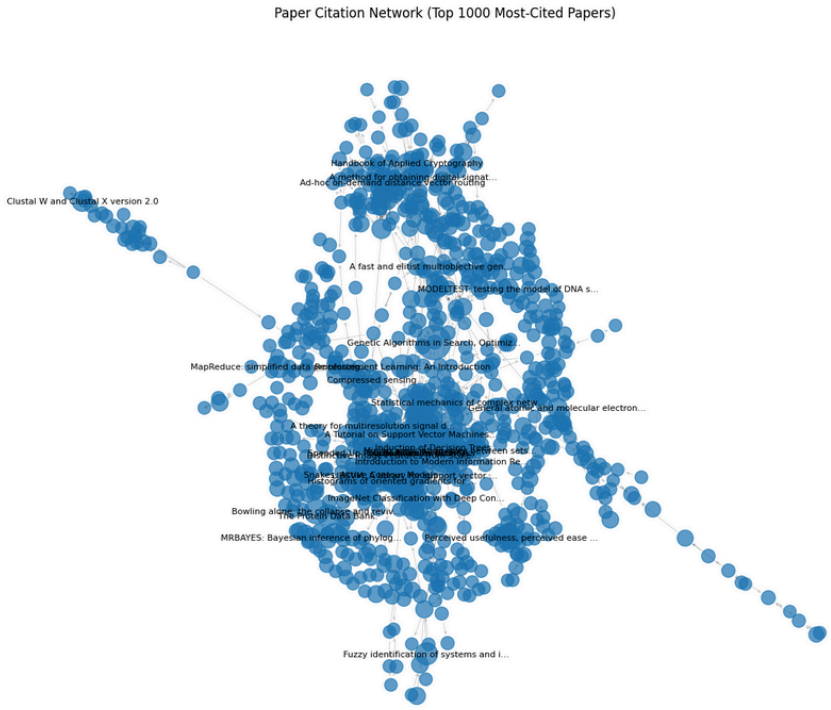
Data Preprocessing

We will use the DBLP-based dataset provided for the class. It contains 3,079,007 papers and 25,166,994 citation links (up to October 2017). Each record has: id, title, authors (list), venue, year, n_citation, references (list of cited paper ids), and abstract. Since this dataset doesn't naturally provide clear numerical features, we took the 'view' where each row is a venue that contains the text of all the titles from that venue, concatenated. In order to do this, we first dropped incomplete rows that have a missing/empty title and venue and removed generic venues such as "arXiv" and "CoRR". We then took only the venues with papers > 1000. We then took each venue and concatenated all the associated titles in said venue into one string. Then on these concatenated titles for each venue, we used the TF-IDF with max_features = 1000 and stop_words = "english" and then applied PCA on the matrix to get it down to 50 dimensions/features and further did z-score scaling on these features.

Exploratory Data Analysis



The author-citation network shows how the top 300 most-cited authors in the DBLP dataset are connected through shared citation relationships. Each node is an author, sized by total citations, and the edges show citation links between them. The dense cluster in the center represents a core group of highly influential researchers whose work is widely cited across multiple areas of computer science. This is where authors like Anil K. Jain (121,850 citations), David E. Goldberg (113,375), and Hari Balakrishnan (109,682) appear, since their research impacts many different subfields. The authors farther from the center have fewer cross-connections, indicating more specialized or narrower citation influence. Overall, the graph highlights how a relatively small group of researchers forms a highly interconnected hub of major contributors.



The paper-citation network shows how the top 1,000 most-cited papers in the dataset are connected through citation links. Each node represents a paper, sized by its total citations, and the dense cluster in the center reflects a core group of highly influential works that are widely cited across many research areas. This includes papers like Genetic Algorithms in Search, Optimization and Machine Learning (73,362 citations), Distinctive Image Features from Scale-Invariant Keypoints (42,508), and Bowling Alone (34,288). Papers on the outskirts tend to be more specialized with a narrower impact. Overall, the network highlights how a small number of landmark papers form the central backbone of citation activity.

The concentration of citations around a small set of highly influential papers mirrors the broader imbalance in the DBLP venue distribution. Because only a few venues dominate the dataset, we focused our classification analysis on the six most frequent venues and balanced them to create a fair evaluation setting.

Venue Selection/Balancing and Dataset Split

To address the extreme class imbalance in the DBLP dataset, where only a few venues dominate the overall publication distribution, we selected the six most frequent venues and balanced the dataset by sampling 3,000 papers from each, producing a selection of 18,000 papers spanning diverse research areas, including LNCS, ICASSP, ICC, ICIP, ICRA, and ISCAS. We split this balanced dataset into 70% training and 30% testing, stratified by venue to ensure equal representation across both sets.

Models

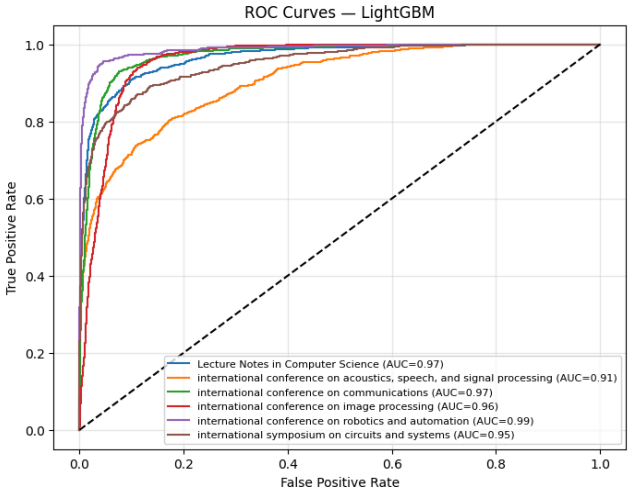
We evaluated six different classification models, all trained on the same 304 dimensional feature vectors (300 SVD text + 4 standardized numeric attributes). These models were chosen to represent several major machine learning families. Parameter sweeps were performed for LightGBM, SVM, and KNN to compare multiple hyperparameter settings and select the best performing configurations. The remaining tree based ensemble models (Random Forest) were evaluated using standard, commonly effective parameter settings.
Overall, LightGBM achieved the strongest performance, followed by Random Forest and SVM. Models like KNN and the single Decision Tree performed substantially worse, likely due to the high dimensionality, sparsity, and nonlinear structure of the feature space. These results highlight that ensemble tree methods are the best suited for this venue classification task.

Training & Testing Time

Model runtimes varied significantly. KNN requires almost no training time because it stores the training data directly, but its prediction time is slower because it must compute distances of all training samples. LightGBM achieved the best balance, with fast training and very fast testing. SVM was by far the slowest in both training and testing because kernel methods scale poorly with high-dimensional data. Random Forest required longer training times due to building many trees, but both predicted quickly. Overall, LightGBM offered the best combination of speed and accuracy.

Model	Train Time (s)	Test Time (s)	Accuracy	Precision	Recall	F1
LightGBM	16.3228	0.1200	0.7969	0.8011	0.7969	0.7954
SVM (RBF)	46.2441	5.4112	0.7544	0.7568	0.7544	0.7534
KNN	0.0453	0.2710	0.4278	0.4467	0.4278	0.4100
Decision Tree	5.3973	0.0020	0.6152	0.6160	0.6152	0.6155
Random Forest	30.2753	0.1398	0.7485	0.7529	0.7485	0.7463

ROC Curve (LGBM)



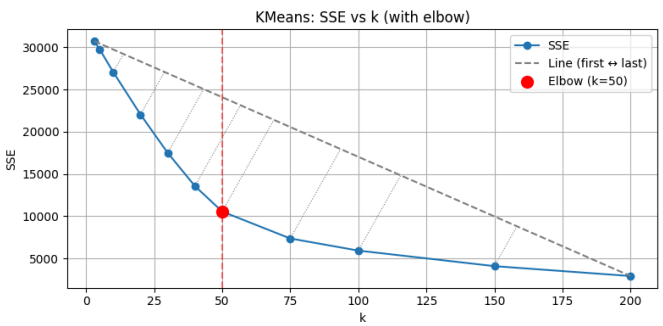
The ROC curves for LightGBM show that it is the strongest model in terms of class separability across all six publication venues. A ROC curve plots the true positive rate against the false positive rate as the decision threshold changes, and the Area Under the Curve (AUC) summarizes how well the model distinguishes one class from the rest, while values closer to 1 indicate stronger discrimination. In our experiment, LightGBM achieves consistently high AUC scores ranging from 0.91 - 0.99, with particular strong performances on ICRA, LNCS, and ICC, all reaching 0.97-0.99. These curves rise sharply towards the top left corner, indicating that LightGBM produces well ranked probability estimates and robust separation between venues even when boundaries are ambiguous. Even the weakest class, ICASSP, still achieves an AUC of 0.91, demonstrating that LightGBM maintains reliable performance across all venue types. Overall, the ROC analysis highlights LightGBM's ability to capture nonlinear structure in the TF-IDF + SVD features and numeric metadata, making it the most effective model in terms of ranking quality

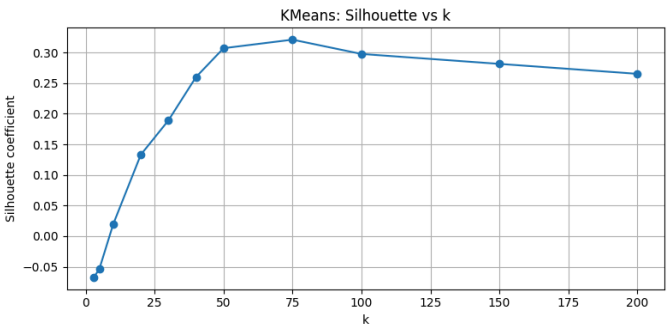
Classification Conclusion

LightGBM delivered the strongest overall performance, achieving 78.7% accuracy with balanced predictions among all six venues. SVD played a key role in reducing dimensionality and improving generalization, while numeric metadata (citations, references, authors, year) contributed additional predictive power. Most misclassifications occur between semantically similar venues such as ICASSP, ICIP, ICC, and ISCAS, indicating that errors are often driven by topical overlap rather than model error. KNN proved unsuitable for high dimensional classification in this setting.

Moving forward with the scaled PCA TF-IDF dataset, we applied three clustering methods (K-means, hierarchical clustering, and DBSCAN) and evaluated them using two main metrics: SSE (sum of squared errors) and the silhouette coefficient.

K-means





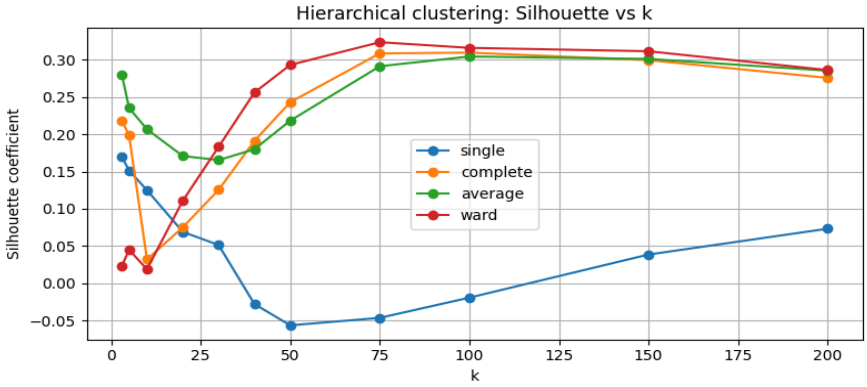
We ran K-means with k in {3, 5, 10, 20, 30, 40, 50, 75, 100, 150, 200} and plotted SSE and silhouette versus k. As expected, SSE decreases steadily as k increases, because more clusters mean shorter distances from each point to its assigned centroid. At the extreme, when k = N (each point is its own cluster), SSE becomes 0.

The silhouette coefficient captures a more meaningful tradeoff by combining within-cluster cohesion and between-cluster separation. At low k, clusters are too broad and heterogeneous, giving low silhouette scores. As k increases, clusters become tighter and better separated, and the silhouette rises until an approximate optimum is reached; beyond that point, increasing k further tends to split coherent groups and gradually decreases the silhouette. In the silhouette-versus-k plot, the score peaks around k = 75.

Using the SSE curve, we also applied the elbow method by looking for where the marginal improvement in SSE begins to diminish. The elbow appears around k = 50, but SSE at k = 75 is still relatively low and close to that elbow value, while the silhouette is slightly higher at k = 75 than at any other tested k. Overall, this suggests that k = 75 offers the best balance between compactness and separation for K-means on this dataset.

Hierarchical clustering

Next, we performed agglomerative hierarchical clustering with four linkage criteria: single, complete, average, and ward. For each linkage type, we cut the resulting dendrogram at the same set of k values used in K-means and computed the silhouette scores.

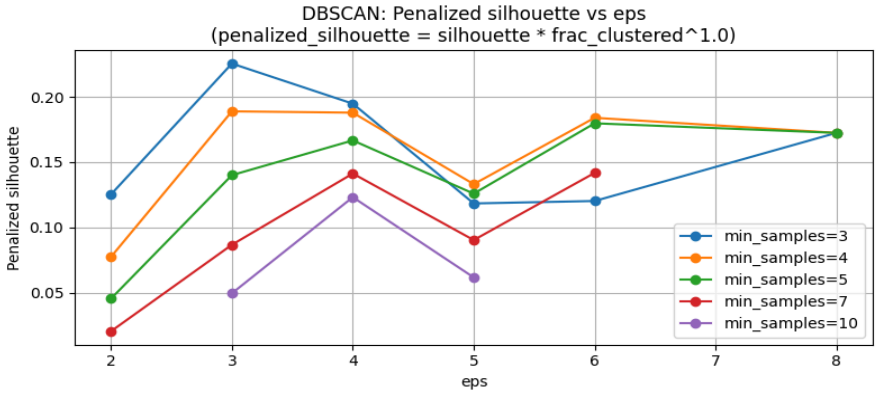


The silhouette-versus-k pattern for hierarchical clustering is broadly similar to K-means: performance improves as k increases up to a certain point and then slowly declines. Again, k = 75 is near-optimal. Among the linkage methods, ward linkage yields the highest silhouette score at k = 75 (0.323), slightly higher than the K-means silhouette at the same k (0.321). Ward linkage explicitly minimizes within-cluster variance at each merge step, which is conceptually similar to minimizing SSE and tends to produce compact, stable clusters. This explains its superior performance relative to single, complete, and average linkage.

DBSCAN

Finally, we explored density-based clustering with DBSCAN, varying eps in {2, 3, 4, 5, 6, 7, 8} and min_samples in {3, 4, 5, 7, 10}. A challenge with using the regular silhouette score for DBSCAN is

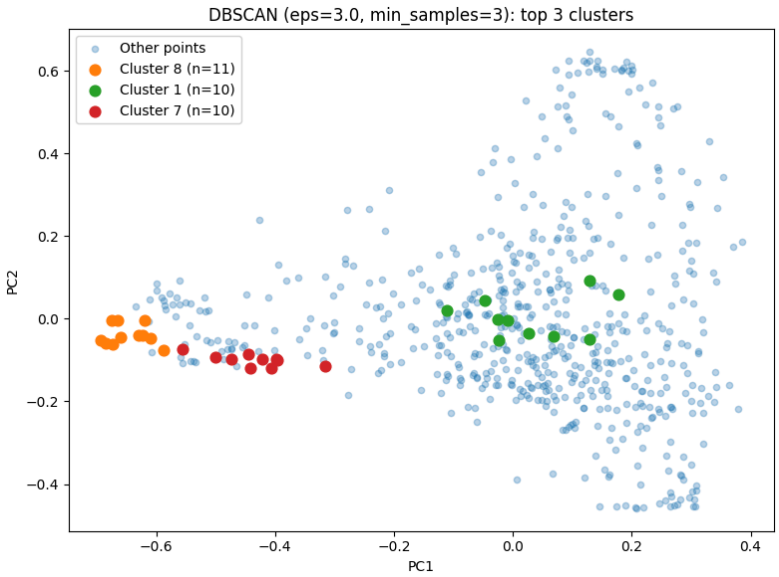
that the algorithm can label a large fraction of points as noise and form a few extremely tight, well separated clusters. This can artificially inflate the silhouette even though most data is ignored. To address this, we defined a penalized silhouette: penalized_silhouette = (silhouette) × frac_clustered, where frac_clustered is the proportion of points assigned to clusters (i.e., not labeled as noise). This discourages solutions that achieve high silhouette at the cost of discarding most data.



From the parameter sweep, eps = 3 and min_samples = 3 gave the best penalized silhouette score (0.225461), producing 44 clusters and 305 noise points. The corresponding unpenalized silhouette is considerably higher (0.432586). This gap likely reflects two factors: (1) DBSCAN's ability to discover non-spherical, arbitrarily shaped clusters that K-means and ward linkage may not capture well, and (2) the removal of borderline or ambiguous points as noise, which improves the average quality of the remaining clusters.

Interpreting the clusters

Using the best DBSCAN configuration (eps = 3, min_samples = 3), we computed a silhouette score for each individual cluster (averaged over its points) and ranked clusters from highest to lowest. we focused on the top three clusters with at least 10 data points, labeled clusters 8, 1, and 7.



The venues in these clusters reveal a clear thematic structure:

- **Cluster 8**: Discrete Applied Mathematics, Discrete Mathematics, Electronic Journal of Combinatorics, Electronic Notes in Discrete Mathematics, European Journal of Combinatorics, etc. This cluster is dominated by discrete mathematics and combinatorics venues.
- **Cluster 1**: ACM SIGSOFT Software Engineering Notes, IEEE Software, IEEE Transactions on Software Engineering, Information & Software Technology, Journal of Systems and Software, and related venues. These are primarily software engineering outlets.
- **Cluster 7**: Applied Mathematics Letters, Applied Mathematics and Computation, Computers & Mathematics with Applications, International Journal of Computer Mathematics, Journal of Computational Physics, and others, representing applied mathematics and computational science.

Each of these clusters is semantically coherent: venues within a cluster share similar topical focus and terminology (discrete math/combinatorics, software engineering, and applied math/computation, respectively). Notably, the discrete math/combinatorics group (Cluster 8) also appears as one of the most clearly defined clusters in both K-means and hierarchical clustering, indicating that this thematic structure is robust across very different clustering algorithms.

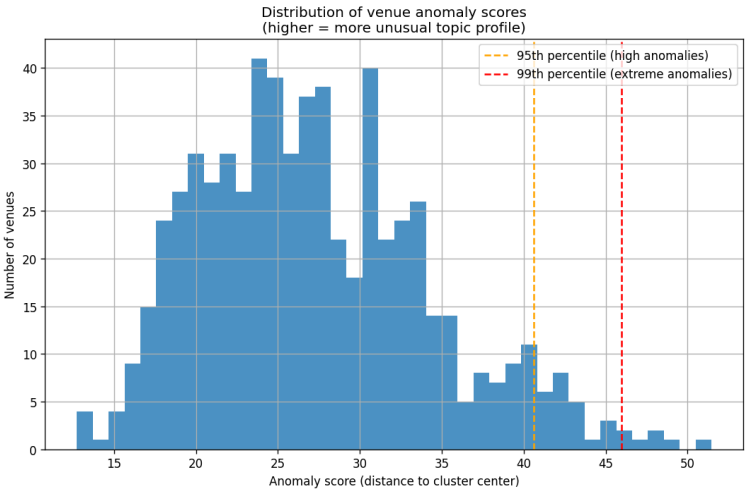
In summary, across K-means, hierarchical clustering with ward linkage, and DBSCAN, the analyses consistently highlight a meaningful partition of the venue space into interpretable research areas, with k around 75 (or a comparable number of density-based clusters) providing a good balance between model complexity and cluster quality.

Anomaly Detection

This clustering work also sets the foundation for the anomaly-detection stage, since identifying what “typical” venue behavior looks like makes it easier to spot the outliers. With that structure in place, we chose an anomaly-detection approach that fits the nature of the DBLP data.

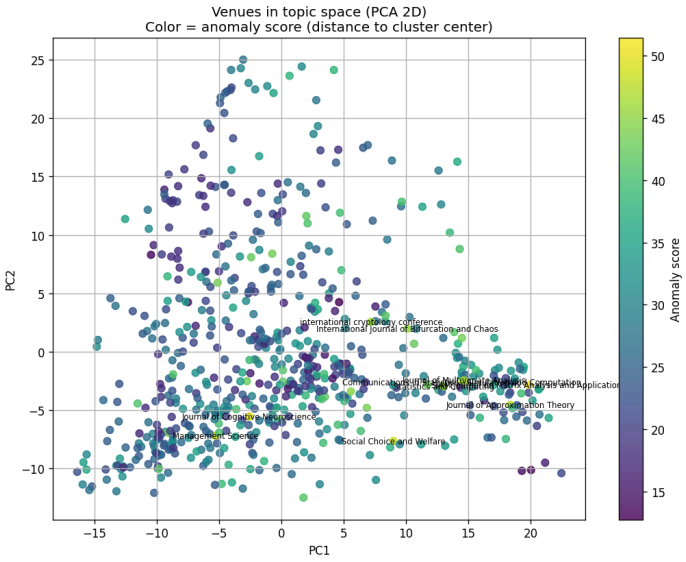
We chose this particular approach because it works well with the kind of data we have. The DBLP dataset is massive, sparse, and high-dimensional, which makes a lot of classical anomaly-detection methods impractical. Methods like k-nearest-neighbors or Local Outlier Factor require enormous amounts of distance computations and behave poorly in high-dimensional TF-IDF space, while DBSCAN does not separate clusters clearly in text-heavy datasets. One-Class SVM also breaks down at this scale. MiniBatch K-Means, on the other hand, handles millions of data points efficiently, and once the clusters are formed, the distance of each venue from its cluster center becomes a simple and intuitive way of measuring how unusual it is. Venues that are far from the center of their topic group naturally represent those with highly specialized or mixed vocabulary, which makes them prime candidates for further analysis.

Interpreting the Histogram of Anomaly Scores



The first figure that we generated is a histogram of the anomaly scores. This plot shows how the venues are distributed across different levels of "unusualness." Scores on the right-hand side represent venues whose topic profiles diverge more strongly from the average for their cluster. we added vertical lines marking the 95th and 99th percentiles so that it is easier to see where the high-anomaly and extreme-anomaly regions begin. This helps define what counts as "unusual" instead of arbitrarily picking a cutoff.

Understanding the PCA Visualization of Venue Outliers



The second figure brings all the venues into a two-dimensional PCA projection and colors each point by its anomaly score. Even though PCA compresses a thousand TF-IDF features into just two dimensions, the plot still provides a good sense of where the outliers sit relative to the rest of the topic map. Venues with higher anomaly scores appear in brighter colors, and they are typically located at the edges of the main cloud or in more isolated regions. We also labeled the top outliers directly on the

scatterplot. This helps connect the numeric anomaly score to real venues in the dataset and makes the visual more meaningful.

What the Highest-Anomaly Venues Reveal About Topic Differences

When we examined the anomaly scores across venues, the ones that stood out were those with vocabularies that don't resemble typical computer science language—mainly very mathematical, interdisciplinary, or niche journals. The clearest example is the SIAM Journal on Matrix Analysis and Applications, which had the highest anomaly score. Compared to the overall TF-IDF profile of all venues—where common terms include "based," "using," "systems," "data," "networks," and "algorithm" SIAM's vocabulary is dominated by tightly focused linear-algebra terms like "matrices," "matrix," "rank," "symmetric," "polynomials," "inverse," and "factorization." Because TF-IDF boosts words that are rare in the global dataset but common within a specific venue, this concentrated mathematical language pushes SIAM far from the cluster centroid. The high anomaly score reflects a real topic separation: SIAM occupies a very different region of the feature space compared to mainstream CS venues.

Conclusion

Across all three components of this project (classification, clustering, and anomaly detection), we found that a unified TF-IDF based representation of publication venues provided a powerful foundation for understanding the structure of the DBLP dataset. The citation network visualization graph showed that influence in computer science was highly concentrated. This imbalance motivated our balanced venue classification method, where LightGBM ended up as the strongest model with high accuracy and AUC values across all six venues since we got the top six. Its performance demonstrated that combining SVD reduced text features with metadata captures meaningful distinctions between research areas, even when venues share overlapping terminology.

Our clustering analyses further revealed that the venue space exhibits coherent and stable topical structure. K Means, Hierarchical Clustering (especially Ward linkage), and DBSCAN all converged towards similar high level grouping, with k=75 offering the best balance between cohesion and separation. The top clusters consistently aligned with interpretable research domains such as discrete math, software engineering, and applied mathematics, indicating that TF-IDF features preserve meaningful semantic differences at the venue level. DBSCAN additionally highlighted smaller dense communities that K means and hierarchical methods tend to smooth over, reinforcing the presence of sharply defined subfields.

Finally, anomaly detection provided a complementary perspective by identifying venues whose vocabularies deviate most strongly from the cluster norms. The venues with the high anomaly scores were typically mathematically intensive or highly specialized journals whose terminology diverges from mainstream computer science language. This supports the idea that distance based anomaly scoring is not only computationally scalable but also semantically interpretable, as the detected outliers correspond to genuine thematic deviation.

Overall, the project shows that combining machine learning models with modern text representation techniques can effectively reveal the structure, boundaries, and outliers of a large dataset. The consistent patterns observed across classification, clustering, and anomaly detection suggest that the DBLP venue landscape contains both well defined research communities and a small number of distinct and specialized outliers, which captures the diversity of the computer science research from the dataset.