Following the Wave or Riding Alone? A Decade of Thematic Evolution at SBBD vs. VLDB

Tiago de Melo[1]

[1]Escola Superior de Tecnologia – Universidade do Estado do Amazonas (UEA)
Manaus – AM – Brazil

tmelo@uea.edu.br

Abstract. *Does SBBD mirror the global database research agenda or cultivate a distinct, locally shaped identity? We study this using large-scale topic modeling of a decade of publications: 489 SBBD and 3,036 VLDB full-text papers from 2016–2025. Applying BERTopic with multilingual embeddings and LLM-assisted label refinement, we extract 31 SBBD and 35 VLDB topics and align them semantically. We observe three patterns: (1) convergence, with 58% of SBBD topics matching VLDB topics, notably in data management, privacy, and NLP; (2) temporal lag, with SBBD adopting mainstream trends 1–4 years after their VLDB peak; and (3) divergence, with 13 SBBD-only topics highlighting a strongly applied, Brazil-centric identity. We conclude with a forward-looking research agenda for SBBD structured around five strategic directions.*

1. Introduction

National and regional scientific venues build shared vocabularies, consolidate local research communities, and often serve as the main publication outlet for researchers limited by language, cost, or institutional support. In the database field, bibliometric studies have examined citation patterns at the Special Interest Group on Management of Data (SIGMOD) and Very Large Data Bases conference (VLDB) [Rahm and Thor 2005], thematic bursts in conference title streams [Kleinberg 2002], and the co-authorship network of SBBD's first 30 editions [Lima et al. 2017]. However, longitudinal comparisons that place the Brazilian Symposium on Databases (*Simpósio Brasileiro de Banco de Dados* - SBBD) in direct thematic dialog with a global reference venue, using full-text topic modeling rather than title-level signals, are still missing from the literature.

SBBD, now in its forty-first edition, is the main scientific event of the Brazilian database community and one of the longest-running computer science symposiums in Latin America. The VLDB, one of the three leading database research venues alongside SIGMOD and International Conference on Data Engineering (ICDE), is used here as a global reference due to its open model and broad scope.

This paper presents, to our knowledge, the first longitudinal, full-text topic-modeling study of SBBD and VLDB over a decade (2016–2025), covering 3,525 papers in two languages. We pair topical analysis with bibliometric data from OpenAlex [Priem et al. 2022] and Semantic Scholar [Kinney et al. 2023]. Our contributions are: (i) a reproducible pipeline to extract, embed, and cluster topics from full-text Portuguese and English proceedings; (ii) a quantitative characterization of the SBBD and VLDB relationship. Our analysis finds 58% topic alignment, a median absolute lag of

$\sim$1.5 years, and 13 SBBD exclusive topics reflecting a distinct applied identity, with privacy research as a notable case where SBBD appears to have preceded the international trend; and (iii) strategic directions for SBBD.

2. Methodology

2.1. Corpus construction and bibliometric data

We collected all publicly available full-text papers from both conferences for 2016–2025. SBBD papers were downloaded from the SBC Open Library (SOL), and VLDB papers from the VLDB site[1]. Text was extracted from PDFs with PyMuPDF [Artifex Software 2024], falling back to Tesseract OCR [Smith 2007] for 31 scanned SBBD 2020 papers.

Table 1 summarizes both corpora[2]. The VLDB corpus has about fifteen times more running words; VLDB papers average 2.4 more pages and twice as many authors. A key feature of SBBD is its bilingualism: 99% of papers mix Portuguese and English, motivating a multilingual embedding model. Bibliometric metadata was collected from OpenAlex (VLDB, 3,167 papers) and Semantic Scholar (SBBD, 488/489 papers).

Table 1. Corpus and bibliometric overview (2016–2025).

Indicator	SBBD	VLDB
Total papers (full-text)	489	3,036
Avg. papers per year (CAGR)	48.9 (+10.5%/yr)	303.6 (+9.4%/yr)
Total running words	2.1M	31.9M
Avg. pages / authors per paper	9.4 / 2.2	12.2 / 4.9
Avg. citations (2016–2022)	2.2	35.1
Avg. references per paper	13.6 (+0.57/yr)	32.6 (+0.21/yr)
% international collab.	21%	38%
% open access	100%	17%
Countries / Unique authors	15 / $\geq$1,111	65 / 8,331

2.2. Topic modeling

We applied BERTopic [Grootendorst 2022] with the multilingual sentence-embedding model `paraphrase-multilingual-MiniLM-L12-v2` [Reimers and Gurevych 2019], which supports the Portuguese–English mix in SBBD papers. UMAP (n_neighbors $= 15$, n_components $= 5$, cosine metric) reduces the embedding space before HDBSCAN clustering (SBBD: min_cluster_size $= 5$, min_samples $= 2$; VLDB: 15, 5). HDBSCAN produced 31 topics for SBBD and 35 for VLDB; 20.4% and 29.2% of documents, respectively, were labeled as outliers and excluded from the temporal analysis. Each topic's top keywords were refined into concise 3–6 word English labels by querying Claude Haiku (Anthropic) with keywords and representative excerpts; all labels were manually reviewed by the author for face validity.[3]

[1]https://www.vldb.org

[2]All study artifacts are available at `https://doi.org/10.5281/zenodo.20763127`.

[3]Claude Haiku was used solely for topic label generation from keyword lists. All labels were verified by the author before use in the semantic alignment step.

2.3. Semantic alignment and lag analysis

We re-encoded all topic labels with the same sentence-transformer and computed a 31×35 cosine-similarity matrix. For each SBBD topic, we selected the best-matching VLDB topic; pairs with label similarity ≥ 0.60 were treated as *matched*. This threshold was chosen empirically: stricter values (e.g., 0.65) reduce the matched set from 18 to 12, dropping well-supported pairs, while looser values inflate coverage with weakly related topics; 0.60 offers the best balance. All 18 matched pairs are corroborated by positive centroid similarity (≥ 0.40), confirming document-level semantic proximity despite cross-lingual confounds. We introduce a simple temporal lag measure for a matched pair (t_S, t_V), defined below, where a positive lag means SBBD peaked *after* VLDB (absorbing an existing trend), while a negative lag means it peaked *before* VLDB (anticipation or independent development).

$$\text{lag}(t_S, t_V) = \text{peak\_year}(t_S) - \text{peak\_year}(t_V)$$

3. Results

3.1. Bibliometric profile

Table 1 reports decade-averaged bibliometric indicators, and Figure 1 shows their annual evolution. Several structural contrasts emerge. **Comparable growth.** Both venues grew about 10% per year, with SBBD increasing from 37 to 91 papers and VLDB from 197 to 443.

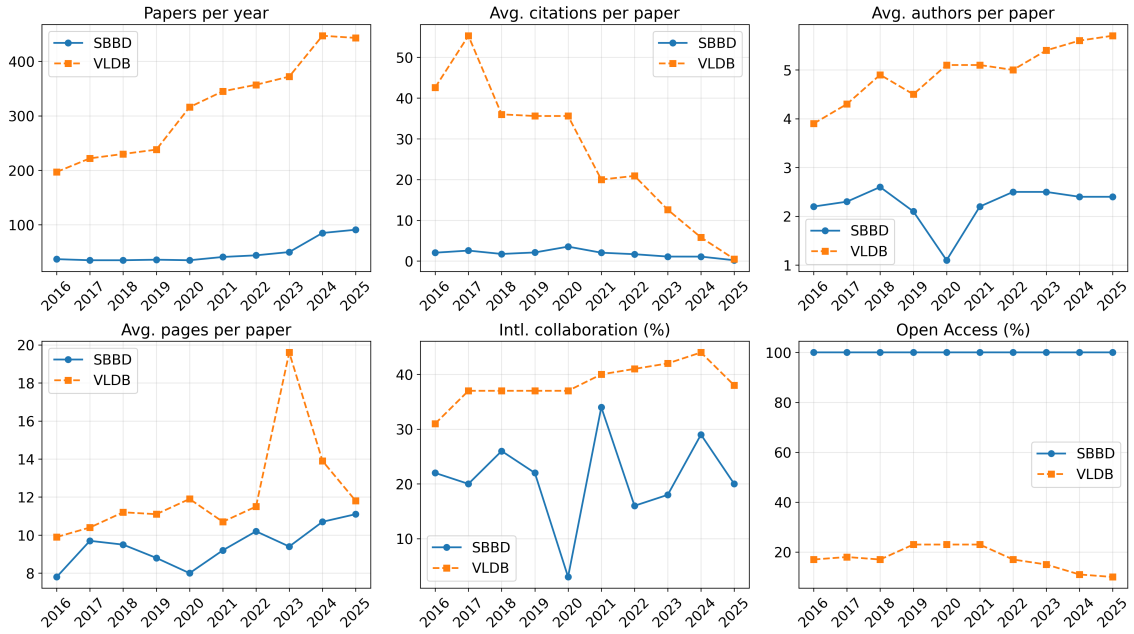

Figure 1. Annual evolution of six bibliometric indicators for SBBD (solid) and VLDB (dashed), 2016–2025.

Smaller teams, shorter papers. SBBD papers average 2.2 authors and 9.4 pages, versus VLDB's 4.9 authors and 12.2 pages, which is typical of a national venue dominated by single-advisor research groups or small-group projects. The 2020 SBBD data point is an outlier in several indicators (1.1 authors/paper, 3% international collaboration, and

a sharp dip in citations/paper; Figure 1). This likely reflects OCR/metadata extraction artifacts from the 31 scanned PDFs of that edition rather than a real structural change: OCR garbles author lists, affiliations, and titles, which also impairs citation matching. The 2020 values should therefore be read with caution across indicators.

Growing international engagement. The average SBBD reference list increased by 0.57 references/year (vs. 0.21 at VLDB), indicating stronger engagement with the broader literature. International collaboration averaged 21% over the decade, peaked at 34% in 2021, then declined slightly.

Open access. SBBD is 100% open access through SOL, a structural advantage over VLDB (whose rate fell from 23% in 2021 to 10% in 2025). Yet citation counts remain far lower (2.2 vs. 35.1 for 2016–2022 papers), likely due to venue prestige, community size, and inconsistent DOI registration.

3.2. Convergence: shared research agenda

Of the 31 SBBD topics, 18 (58%) matched at least one VLDB topic above the 0.60 threshold. Table 2 lists the top-8 similarity pairs. The strongest alignments are in core database areas, such as data management, privacy, and query optimization, showing a common disciplinary backbone.

Table 2. Top-8 matched topic pairs (SBBD $\leftrightarrow$ VLDB) by semantic similarity.

SBBD topic	VLDB topic	Sim.
Data Models and Dataset Management	Dataset Management and Data Quality	0.84
Data Management and Governance	Data Management and Processing Systems	0.81
Privacy-Preserving Data Management Systems	Differential Privacy for Database Protection	0.81
Hardware-Aware Database Query Optimization	Query Optimization and Processing Techniques	0.78
Graph Processing and Database Performance	Main-Memory Database Performance Optimization	0.77
Data Source Quality and Profiling	Dataset Management and Data Quality	0.75
Data Quality and Entity Resolution	Dataset Management and Data Quality	0.71
Graph-Based Data Organization and Management	Graph Clustering and Network Analysis	0.71

A notable convergence is the rise of NLP-oriented research in both venues. In SBBD, *Text Classification and NLP* are the fastest-growing topics ($+1.47$ p.p./year), while at VLDB, *Natural Language Querying of Databases* grows at $+0.42$ p.p./year. Both conferences are converging on the intersection of LLMs and data systems.

3.3. Temporal lag: how fast does SBBD absorb trends?

Table 3 summarizes the lags for matched pairs with $|\text{lag}| \geq 1$ year (7 of the 18 matched pairs have zero lag and are omitted), ranging from -3 to $+4$ years (median absolute lag across all 18 pairs: 1.5 years). Given the relatively small SBBD corpus ($\sim$49 papers/year), peak-year estimates for individual topics are sensitive to sampling variation and should be read as indicative rather than definitive. Privacy research peaks earlier in SBBD than in VLDB (lag $= -3$). This peak falls in the OCR-affected 2020 edition, but topic assignment relies on distinctive lexical markers (e.g., LGPD, anonymization) across the full text, making it far less sensitive to OCR noise than the bibliometric indicators above. A plausible driver is COVID-19, which spurred 2020 demand for privacy-preserving methods on sensitive health data, absorbed faster by SBBD's shorter editorial cycle.

Table 3. Temporal lag (in years) for matched pairs with non-zero lag.

SBBD topic	VLDB topic	Lag	Sim.
Database Applications in Medical and Scientific Data	Data Analysis and Query Processing	$+4$	0.63
Data Management and Governance	Data Management and Processing Systems	$+3$	0.81
Similarity Join Algorithms and Data Integration	Set Similarity Search and Matching	$+3$	0.69
Web Data Integration and Quality	DB Query Processing and Data Integration	$+3$	0.65
Text Classification and NLP	Natural Language Querying of Databases	$+1$	0.65
Hardware-Aware DB Query Optimization	Query Optimization and Processing Techniques	$+1$	0.78
RDF Data Storage and Query Processing	DB Data Processing and Schema Management	$+1$	0.67
Data Models and Dataset Management	Dataset Management and Data Quality	-1	0.84
Data Analytics for Brazilian Applications	Data Analytics and Recommendation Systems	-1	0.63
Graph Processing and Database Performance	Main-Memory DB Performance Optimization	-1	0.77
Privacy-Preserving Data Mgmt.	Differential Privacy for DB Protection	-3	0.81

3.4. Divergence: what makes SBBD unique?

The remaining 13 SBBD topics (42%) have no semantic counterpart in VLDB and fall into two categories. **Domain-applied, Brazil-centric** topics address locally relevant challenges: *Public Transportation Data Management*, *Urban Traffic Anomaly Detection*, *Time Series Forecasting for Brazilian Environmental Applications*, and *Data Ecosystems and Economic Analysis*. **Methodological niches** include *Scientific Workflow Data Processing Optimization*, *Ontology Alignment and Topic Modeling*, and *Spatial and Geographic Data Management*. Symmetrically, 10 globally prominent VLDB topics are absent from SBBD, notably *Cloud Database Systems* and *Machine Learning on Distributed Databases*.

4. Discussion and Vision

The three phenomena, convergence, temporal lag, and divergence, show a venue that has absorbed the global disciplinary core while building a research identity distinct from a time-shifted copy of VLDB. This identity reflects a strong applied focus in nationally relevant areas, an evolving regulatory landscape that creates unique research opportunities, and a community of small, single-advisor research groups. We translate these findings into a forward-looking agenda of five strategic directions, each a concrete opportunity for the community to act before the international agenda consolidates.

The systems research gap. The near-absence of large-scale systems research at SBBD (graph analytics, cloud-native databases, distributed ML pipelines) likely reflects the long implementation cycles and industrial partnerships that such work demands. This gap is also an opportunity for contributions on heterogeneous and resource-constrained environments. Edge data management and systems for intermittent connectivity are underserved problems with real methodological depth.

LLMs and database systems. Text Classification and NLP is SBBD's fastest-growing topic, mirroring VLDB's trajectory. SBBD risks remaining only a consumer of LLM technology for database tasks. Alternatively, it can focus on genuinely Brazilian problems: privacy-preserving text-to-SQL under LGPD, differential privacy for fine-tuning on sensitive government data, and retrieval-augmented generation over heterogeneous public databases. Brazil's strong NLP ecosystem, exemplified by widely adopted Brazilian Portuguese models such as BERTimbau [Souza et al. 2020], combined with database systems expertise, is an underused asset.

Regulatory opportunities. Brazil's data-protection framework (LGPD) and its forthcoming AI bill (PL 2338/2023) are creating regulatory pressure on AI-integrated data systems ahead of many other jurisdictions. This regulatory context offers SBBD a distinctive research niche—compliance-aware query processing, auditable data pipelines, and privacy-by-design for public-sector databases, that is not yet well served by the international community.

Applied research is a strength. Work on public transportation, urban traffic, environmental time series, and electoral data is not less rigorous than database research; it is more challenging due to data quality, schema heterogeneity, and governance constraints of Brazilian public-sector data. For example, environmental monitoring in the Amazon basin produces data at a scale and with spatio-temporal complexity that challenges stream processing, spatial indexing, and anomaly detection beyond what standard benchmarks cover. Positioning SBBD where Brazilian public data challenges drive advances in database methodology would both distinguish the symposium internationally and foster a virtuous cycle between researchers and the institutions that manage these data assets.

Open science. SBBD's 100% open-access rate is a structural advantage that is underused: most papers lack consistent DOIs, are not promptly indexed in major citation databases, and are rarely extended into international journals. Treating SBBD papers as the first step in an international dissemination chain, rather than the final destination, would amplify their reach. In this spirit, we make all artifacts from this study publicly available: the per-paper bibliometric dataset, the topic model outputs (topic assignments, temporal evolution tables, and Claude-refined labels for both corpora), and the full analysis pipeline.

Priorities and urgency. Of these five directions, *LLM–database integration* and *regulatory opportunities* are time-sensitive. Both are arenas where the international agenda is moving fast; SBBD has a narrow window to shape, not merely follow. The systems research gap and open science infrastructure are longer-horizon investments that require sustained community effort. Applied Brazilian research is already the community's strongest differentiator and should be explicitly positioned as such in calls for papers, special tracks, and international outreach. Together, these five directions define a tractable agenda for the next decade of SBBD.

5. Limitations

HDBSCAN labels 20.4% (SBBD) and 29.2% (VLDB) of documents as outliers, mainly papers spanning multiple topics rather than missing themes. The 0.60 cosine similarity threshold is an empirical design choice. Stricter thresholds reduce matches from 18 to 12, while looser ones add weakly related pairs. The peak-year lag metric can be misleading for flat topic distributions, and comparing SBBD only with VLDB limits generalizability. Finally, using full text instead of abstracts maximizes topical signal but may slightly overemphasize technical topics from methodological sections in both venues.

6. Conclusion

We conducted a decade-long, full-text topic-modeling study of SBBD and VLDB (2016–2025), covering 3,525 papers in two languages. SBBD matches VLDB on 58% of topics,

adopts trends with a median absolute lag of $\sim$1.5 years, and preserves a distinct applied profile through 13 topics absent in VLDB. Looking ahead, SBBD's key strategic choices are to invest in systems research, leverage the NLP–database convergence, and position its applied Brazilian research agenda as a strength.

Acknowledgments

The author acknowledges the support provided by the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT–TILD–IAR), funded by the Brazilian National Council for Scientific and Technological Development (CNPq), grant no. 408490/2024-1.

References

Artifex Software (2024). PyMuPDF: Python bindings for MuPDF.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Kinney, R., Anastasiades, C., Arthur, R., et al. (2023). The Semantic Scholar Open Data Platform. *arXiv preprint arXiv:2301.10140*.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 91–101.

Lima, L. H. C., Penha, G., Rocha, L. M. d. A., Moro, M. M., da Silva, A. P. C., Laender, A. H. F., and de Oliveira, J. P. M. (2017). The collaboration network of the Brazilian Symposium on Databases. *Journal of the Brazilian Computer Society*, 23(1):10.

Priem, J., Piwowar, H., and Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

Rahm, E. and Thor, A. (2005). Citation analysis of database publications. *ACM SIGMOD Record*, 34(4):48–53.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Smith, R. (2007). An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.