

WEAVENET: INTERACTIVE HIERARCHICAL CUSTOM BUILT GRAPH DATABASE

Presented by:

Anvay Bhanap
Chetana Muralidharan
Kedar Prashant Vichare
Meera Alpesh Dedhia
Mounashree Prasanna

Project Advisors:

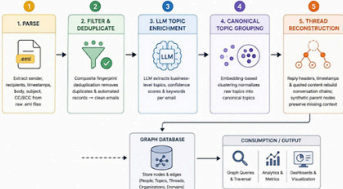
Dr. Simon Shim

Introduction

Enterprise email archives hold vast organizational knowledge - yet traditional keyword search fails to reveal *who* collaborates with *whom*, which topics dominate, and how people are connected across the organization. **WeaveNet** transforms the Enron email corpus - 517,401 raw emails from a real corporate network - into a semantically rich, interactive graph database. The system supports communication-pattern discovery, compliance analysis, and organizational intelligence through graph traversal, approximate search, and a React-based visual explorer - going far beyond what keyword retrieval alone can offer.

Methodology

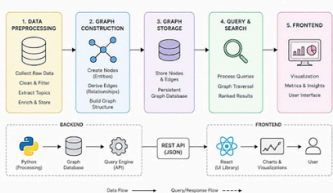
Data Pipeline:



Graph Statistics:

Total Nodes	764,737
Total Edges	2,860,219

System Architecture:



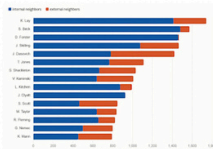
Algorithm Families:

Family	Algorithms
Neighbor Traversal	HNSW vs. NSG
Path Discovery	ScaNN vs. IVF-Flat
Content/Keyword Search	LSH vs. SimHash
Ranking & Similarity	Annoy vs. VP-Tree
Direct Index Lookup	Index-Only

Analysis and Results

Analytics:

Bridge node analysis: Each bar shows internal (enron.com) vs external communication partners. 3,316 persons act as bridges; only top 15 shown.



Algorithm Performance Summary:

Algorithm Pair	Key Metric	Winner Score	Runner-Up Score
HNSW vs NSG	Recall@K	0.943	0.729
HNSW vs NSG	Edge-weight ordering accuracy	0.939	0.840
ScaNN vs IVF-Flat	Path success rate	0.908	0.790
ScaNN vs IVF-Flat	Path completeness	0.949	0.887
LSH vs SimHash	Precision@K	0.858	0.813
LSH vs SimHash	False positive rate	0.142	0.187
Annoy vs VP-Tree	NDCG@K	0.847	0.822
Annoy vs VP-Tree	Recall@K	0.895	0.863
Index Lookup	Hit rate / Sort correctness	1.000	—

Query Latency by Family:

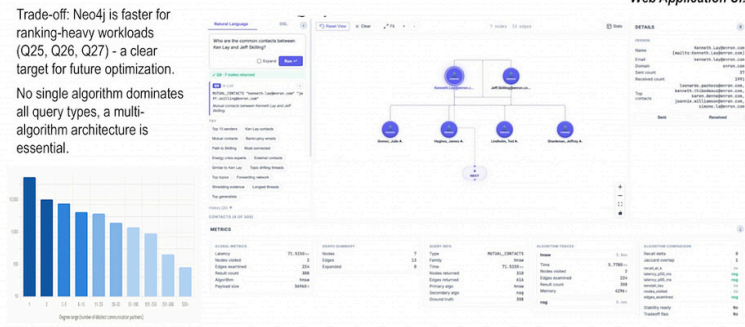
Query Family	Avg. Latency	Suitable For
Index Lookup	7.3 ms	Real-time interactive lookup
HNSW Neighbor Traversal	16.4 ms	Interactive graph expansion
LSH Content Search	113 ms	Analytical content search
ScaNN Path Discovery	273 ms	Constrained path analysis
Annoy Ranking	287 ms	Top-K ranking & similarity

Neo4j Benchmark Highlights:

Query	Our Engine (p50)	Neo4j (p50)	Speed-up
Neighbor lookup	0.548 ms	3.440 ms	6.3x
Path traversal	0.070 ms	5.112 ms	73x
Keyword search	4.360 ms	53.597 ms	12.3x
Direct lookup	0.063 ms	2.877 ms	45.7x
Index lookup	0.072 ms	5.676 ms	79.4x

Trade-off: Neo4j is faster for ranking-heavy workloads (Q25, Q26, Q27) - a clear target for future optimization.

No single algorithm dominates all query types; a multi-algorithm architecture is essential.



Real-world Use Cases:

- Enterprise Search:** Retrieve emails by topic, person, or relationship, not just keywords
- Compliance & Legal Discovery:** Trace communication between people, orgs, and time periods
- Organizational Network Analysis:** Find communication hubs, collaboration clusters, and info flow patterns
- AI Assistant Grounding:** Power evidence-backed answers: "Who discussed this agreement?"

Summary/Conclusions

WeaveNet successfully transforms a noisy corporate email corpus into a context-aware, semantically enriched graph intelligence system. No single algorithm dominated all query types - HNSW excels at neighbor traversal, ScaNN at path discovery, LSH at content search, and Annoy at ranking - confirming the value of a multi-algorithm architecture. Direct index lookups outperform all approximate methods for exact queries and are significantly faster than Neo4j for those workloads.

Key References

- [1] B. Klimt and Y. Yang, "The Enron Corpus: A New Dataset for Email Classification Research," *ECML*, 2004.
- [2] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
- [3] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [4] Y. A. Malkov and D. A. Yashunin, "Efficient and Robust Approximate Nearest Neighbor Search Using HNSW," *IEEE TPAMI*, 2018.
- [5] A. Gionis, P. Indyk, and R. Motwani, "Similarity Search in High Dimensions via Hashing," *Vldb*, 1999.
- [6] M. S. Charikar, "Similarity Estimation Techniques from Rounding Algorithms," *STOC*, 2002.

Acknowledgements

We would like to thank our Project Advisor, Simon Shim, for their continued guidance, feedback, and support throughout the development of this project.

We also extend our gratitude to the Applied Data Science Department for providing the resources, framework, and opportunity to undertake this work.

Interested in partnering with us on a capstone project or exploring other ways to collaborate with the College of IDS? Contact CIDS_CAP@sjsu.edu to discuss potential projects and partnership opportunities.