technische universität dortmund

sisap

Theoretical and Practical Insights into Graph-Based Indexing

Erik Thordsen and Erich Schubert

TU Dortmund, Informatik 8 Data Mining

{firstname.lastname}@tu-dortmund.de

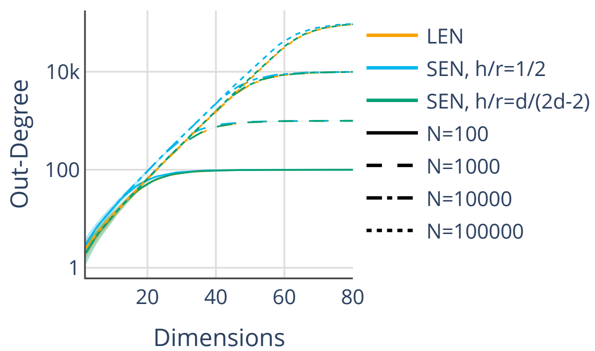
SISAP 2025

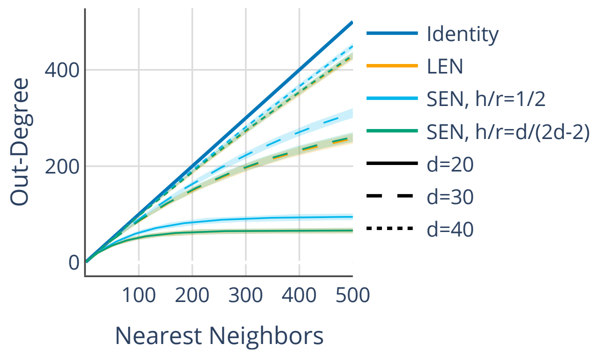
Motivation

- Graph-based indexing (HNSW, DEG, etc.) works well, but why?
- Literature claims quality based on graph classes (RNG/LEN, SSG/SEN, etc.)
- In dense, high-dim. data, we argue that recall rather follows probabilistics
- The long-term goal is better understanding and autoparametric approaches
- For experiments, new competitive Rust libraries

Graph Classes

- Out-degree of RNG and SSG grows with $\Gamma(d)$
- Practical sparse graphs cannot approximate that
- For small enough k, kNNG $\subsetneq$ RNG, SSG
- Recall must follow from other properties



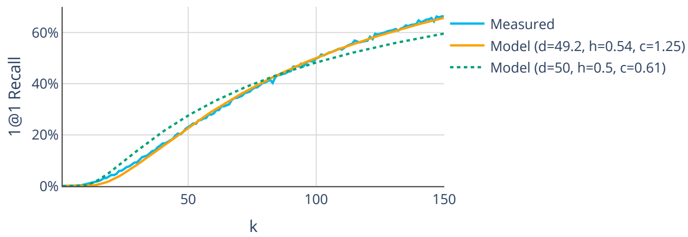


Analysis of Capped Beam Search on kNNG

- We cap the back tracking queue of beam search to size Q
- Probability of a single vertex having a neighbor towards q from geometry
- Probability of the queue having such a neighbor from probability amplification
- Probability of finding the 1-nn from repeated trials with expected path length L

$$p_{beam}(\text{success}) = \left(1 - \left(1 - k/\mathbb{E}^{(N)}_{sector}\right)^{Q}\right)^{L}$$

N is the dataset size, $\mathbb{E}^{(N)}_{sector}$ the expected number of sector exclusion neighbors in N points



This simple model fits real recall okay, but better understanding of L needed.

Implications

- Approximating RNG/SSG in high-dim. is futile
- kNNG recall can likely (for synthetic data) be modeled
- Automatic suggestion for parameter choice given desired recall likely possible
- Multiple immediate results, e.g.:
  - Capped beam search theoretically useful and practically faster
  - Random edges can improve recall
  - Hierarchies in high-dim. much less useful

* Source code at https://github.com/eth42/GraphIndexAPI/
https://github.com/eth42/GraphIndexBaselines/