Overview

My name is Kexin Rong. I am a Ph.D. student at Stanford University, co-advised by Peter Bailis and Philip Levis.

I build systems and design algorithms that use synopses and sampling techniques to prioritize computation over inputs that have the most impact on downstream analytics tasks.

My projects are inspired by real-world analytics tasks, including:

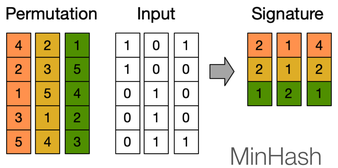
- Scientific analysis (this poster):
  - Earthquake Detection [VLDB'18]
  - Density Estimation [ICML'19]
- Infrastructure monitoring:
  - Anomaly detection and explanation [SIGMOD'17]
  - Time series visualization [VLDB'17]
- Analytical queries in big-data clusters
  - Partition-level sampling [VLDB'20]

Background: Locality Sensitive Hashing

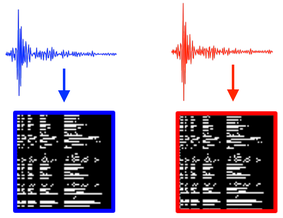
This poster presents two results from one area of my work, prioritizing computation via novel uses of Locality Sensitive Hashing (LSH).

LSH hashes similar inputs to the same "bucket" with high probability.

Similarity Search with LSH



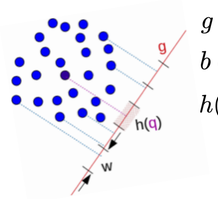
MinHash



Collision prob ~ Jaccard similarity

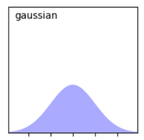
$$\mathbb{P}[h(A) = h(B)] = J(A, B)$$

Importance Sampling with LSH



$$g \sim \mathcal{N}(0, I_d)$$
$$b \sim U[0, 1]$$
$$h(q) = \lceil \frac{g^\intercal q}{w} + b \rceil$$

Euclidean LSH



closer points "weight" more

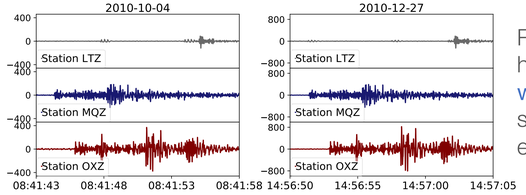
difficult case for random sampling

Collision prob ~ Sampling distribution

$$\mathbb{P}[h^\kappa(x) = h^\kappa(q)] \propto k(x, q)$$

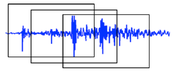
Earthquake Detection

Goal: Detect Unknown Small Earthquakes



Repeating earthquakes have near identical waveforms at the same station, regardless of the earthquake's magnitude.

1 *Feature Extraction*
Input: waveform
Output: binary fingerprints



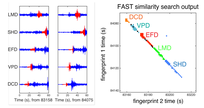
waveform

2 *Similarity Search via LSH*
Input: binary fingerprints
Output: sparse similarity matrix

fingerprints

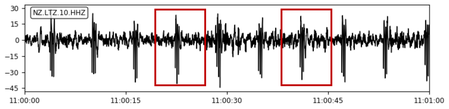
similarity matrix

3 *Filter and Alignments*
Input: each station's similarity matrix
Output: potential earthquakes

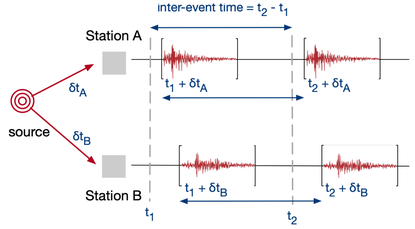


earthquake candidates

Optimizations: Leveraging Domain Priors



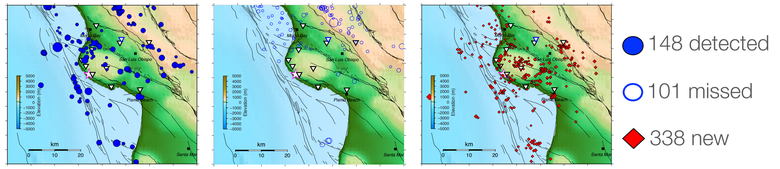
Prior: Earthquake Frequency
Effect: Exclude repeating noise patterns from similarity search



Prior: Propagation time
Effect: Cross-reference results from different seismic stations to reduce false positives

Case Study: Diablo Canyon nuclear power plant

Data: 11 stations, 27 channels, up to 10 years of data
Performance: >100x speedups end-to-end
Result: 3957 catalog earthquakes, 597 new local events



148 detected
101 missed
338 new

Kernel Density Estimation

Goal: $(1 \pm \epsilon)$-approximation to $Z_\chi(q)$ for query $y \in \mathbb{R}^d$

- Dataset: $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$
- Kernel Density for query $q \in \mathbb{R}^d$: $Z_\chi(q) = \sum_{i=1}^{n} \mu_i \cdot k(x_i, q)$

1 *Sketch the dataset: Reduce preprocessing overhead*

HBS: hashing + non-uniform sampling

for $j = 1, \ldots,$ SketchSize:
- Sample bucket i prob. $\propto {n_i}^\gamma$
- Sample a random point J from bucket i: $S \leftarrow S \cup \{J\}$
- Weight it so that $\mathbb{E}_J[\hat{w}_J k(q, x_J)] \propto \text{KDF}_P(q)$

return $(\hat{w}, S)$

Theorem: $O(1/\tau)$ points suffice.
- Approx. any density $\mu \geq \tau$.
- Reduce space from $O(n/\sqrt{\tau})$ to $O(1/\sqrt{\tau^3})$
- Contains a point from any bucket with $\geq n \cdot \tau$ points

2 *Diagnosis: Estimate dataset-specific performance*

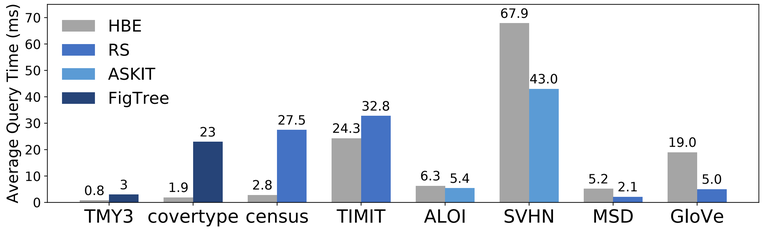
Variance bound for beyond-worst case scenarios

Average relative variance ∝ sample complexity: $\mathbb{E}_{q \sim P}[\frac{\mathbb{V}[Z(q)]}{\mathbb{E}[Z(q)]^2}]$

3 *Adaptive Sampling: Stop early for easier queries*

for $t = 1, \ldots, T$ do — `# for each level`
- $\mu_t \leftarrow 2^{-t}$ — `# current guess of the mean`
- $m_t \leftarrow \lceil \frac{6}{\epsilon^2} V_t(\mu_{t+1}) \rceil$ — `# sufficient sample size in level t`
- $Z_t^{(i)} \sim \mathcal{Z}(t, y)$ — `# get i.i.d. unbiased estimates`
- $\hat{Z}_t \leftarrow mean\{Z_t^{(1)}, \ldots, Z_t^{(m_t)}\}$
- if $\hat{Z}_t \geq \mu_t$: return $\hat{Z}_t$ — `# consistency check`

Real-world Datasets: up to 10x speedup over second best



Diagnosis: correctly predicts all except the SVHN dataset