Set Similarity Search Beyond MinHash

Tobias Christiani Rasmus Pagh
IT University of Copenhagen

SCALABLE SIMILARITY SEARCH
erc European Research Council
Established by the European Commission

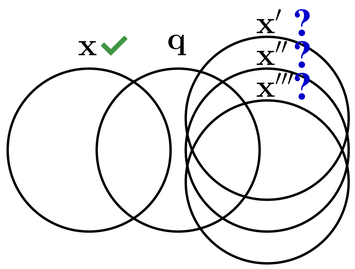
Problem

- Set similarity search:
  - Preprocess a collection P of n subsets of $[d] = \{1, \ldots, d\}$.
  - Given a query set $\mathbf{q} \subseteq [d]$ return $\mathbf{x} \in P$ such that $\mathrm{sim}(\mathbf{q}, \mathbf{x}) \geq s_1$.
- Simplification: All sets have size t.
  - $\mathrm{sim}(\mathbf{x}, \mathbf{y}) = |\mathbf{x} \cap \mathbf{y}|/t$.
- Approximation: Allow returning $\mathbf{x} \in P$ with $s_1 > \mathrm{sim}(\mathbf{q}, \mathbf{x}) > s_2$.

$\mathrm{sim}(\mathbf{q}, \mathbf{x}) \geq s_1$ ✓

$s_1 > \mathrm{sim}(\mathbf{q}, \mathbf{x}) > s_2$?

$\mathrm{sim}(\mathbf{q}, \mathbf{x}) \leq s_2$ ✗



Contribution

Data structure solution
Space: $n^{1+\rho} + tn$ Query time: tn^{ρ}

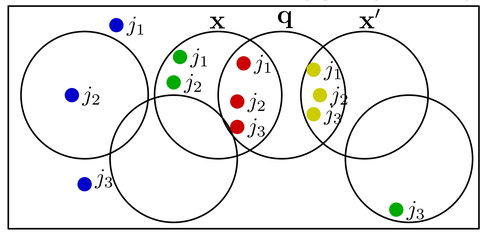
Upper bound	Lower bound*
$\rho = \frac{\log(1/s_1)}{\log(1/s_2)}$	$\rho \geq \frac{\log(1/s_1)}{\log(1/s_2)} - o_d(1)$

- Example: $s_1 = 0.5, s_2 = 0.25$ gives $\rho = 0.5$
  - MinHash $\rho = 0.56$
- Solution for Braun-Blanquet similarity:

$$\mathrm{sim}(\mathbf{x}, \mathbf{y}) = \frac{|\mathbf{x} \cap \mathbf{y}|}{\max(|\mathbf{x}|, |\mathbf{y}|}$$

Random buckets

Collection P of sets $[d] = \{1, \ldots, d\}$



- Collection B of random buckets $(j_1, \ldots, j_k) \in [d]^k$.

$$B(\mathbf{x}) = \{(j_1, \ldots, j_k) \in B \mid j_1 \in \mathbf{x} \wedge \cdots \wedge j_k \in \mathbf{x}\}$$

Query time: $\underbrace{|B(\mathbf{q})|}_{\text{Lookups}} + \underbrace{\sum_{\mathbf{x} \in P:\ \mathrm{sim}(\mathbf{q},\mathbf{x}) \leq s_2} |B(\mathbf{q}) \cap B(\mathbf{x})|}_{\text{Collisions}} + \underbrace{|B|}_{\text{Buckets}}$

Analysis

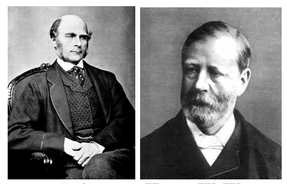
- Correctness: $\mathbf{q}, \mathbf{x}$ with $\mathrm{sim}(\mathbf{q}, \mathbf{x}) \geq s_1$.
  - Pr. collision $(|\mathbf{q} \cap \mathbf{x}|/d)^k \geq (s_1 t/d)^k$.
  - Use $b = \lceil (s_1 t/d)^{-k} \rceil$ buckets.
- Lookups:
  - $E[|B(\mathbf{q})|] = b(t/d)^k = (1/s_1)^k$.
- Comparisons: $\mathbf{q}, \mathbf{x}'$ with $\mathrm{sim}(\mathbf{q}, \mathbf{x}') \leq s_2$.
  - $E[|B(\mathbf{q}) \cap M(\mathbf{x}')|] = b(|\mathbf{q} \cap \mathbf{x}'|/d)^k \leq b(s_2 t/d)^k = (s_2/s_1)^k$.
- Balance: $k = \log(n)/\log(1/s_2)$
  - Space: $n^{1+\rho} + tn$.
  - Query time: tn^{ρ}

$|B| \gg n \gg |B(\mathbf{q})|$

Branching processes

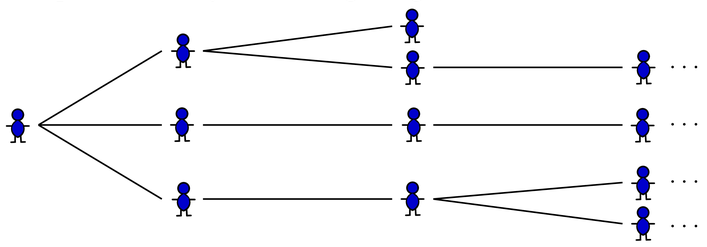
- Study of population growth in the 1870's.
  - Will my surname survive k generations?
- Offspring X i.i.d. across generations.
  - Example: $X \sim \mathrm{bin}(3, 1/3)$

Lemma [Agresti 1974] Let $\mathbb{E}[X] = 1$, then the the probability of surviving k generations is at least $1/(k\mathrm{Var}[X] + 1)$.



Francis Galton (1822-1911) Henry W. Watson (1827-1903)

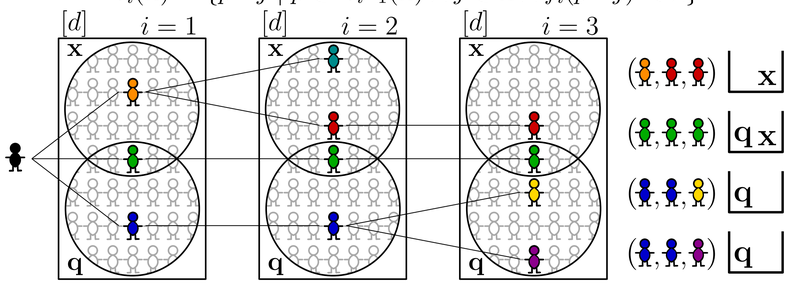
- Example: Probability of surviving 100 generations is at least 0.01.



Chosen Path

- Buckets $B(\mathbf{x})$ as paths $p = (j_1, \ldots, j_k)$ in a branching process.
- For $i = 1, \ldots, k$ sample random hash functions $f_i \colon [d]^i \to [0, 1]$.

$$B_1(\mathbf{x}) = \{j \mid j \in \mathbf{x} \wedge f_1(j) < \lambda\}$$
$$B_i(\mathbf{x}) = \{p \circ j \mid p \in B_{i-1}(\mathbf{x}) \wedge j \in \mathbf{x} \wedge f_i(p \circ j) < \lambda\}$$



- Correctness: Offspring $X = |B_1(\mathbf{q}) \cap B_1(\mathbf{x})| \sim \mathrm{bin}(|\mathbf{x} \cap \mathbf{q}|, \lambda)$.
  - Set $\lambda = 1/s_1 t$: For $\mathrm{sim}(\mathbf{q}, \mathbf{x}) \geq s_1$ we have $\mathbb{E}[X] \geq 1$.

Compute $B(\mathbf{x})$ in time $O(|B(\mathbf{x})|)$!!!

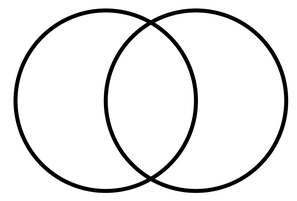
Beyond MinHash?

- Probability $p(\mathrm{sim}(\mathbf{q}, \mathbf{x}))$ of sets colliding in bucket.
  - Performance $\rho = \log(1/p(s_1))/\log(1/p(s_2))$.

MinHash [B '97]

$h(\mathbf{q}) = \arg\min_{j \in \mathbf{q}} f(j)$

Random $f \colon [d] \to [0, 1]$.

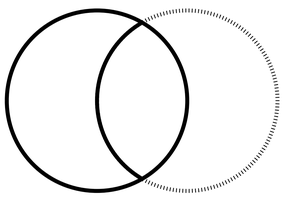


$|\mathbf{q} \cap \mathbf{x}|/|\mathbf{q} \cup \mathbf{x}|$

Chosen Path

$B(\mathbf{q}) = \{j \in \mathbf{q} \mid f(j) < \lambda\}$

Random $f \colon [d] \to [0, 1]$.



$|\mathbf{q} \cap \mathbf{x}|/|\mathbf{q}|$

Previous work

- Existing solutions:
  - Space $n^{1+\rho} + tn$.
  - Query time: tn^{ρ}.
- Comparison:
  - Blue: $\rho_{\text{new}} < \rho_{\text{other}}$
  - Red: $\rho_{\text{new}} \geq \rho_{\text{other}}$
- Observations:
  - $\rho_{\text{new}} < \rho_{\text{minhash}}$
  - $\rho_{\text{new}} < \rho_{\text{angular}}$
  - $\rho_{\text{new}} \lesssim \rho_{\text{data}}$

Two regimes:

- Data-dependent:
  - $(s_1, s_2, P) \to \mathcal{M}$.
- Data-independent:
  - $(s_1, s_2, |P|) \to \mathcal{M}$.

Reference	ρ
This paper	$\log(1/s_1)/\log(1/s_2)$
MinHash [B'97 + IM'98]	$\log \frac{s_1}{2-s_1} / \log \frac{s_2}{2-s_2}$
Angular LSH [AILRS'15]	$\frac{1-s_1}{1+s_1} / \frac{1-s_2}{1+s_2}$
Data-dep. LSH [AR'15]	$\frac{1-s_1}{1+s_1-2s_2}$

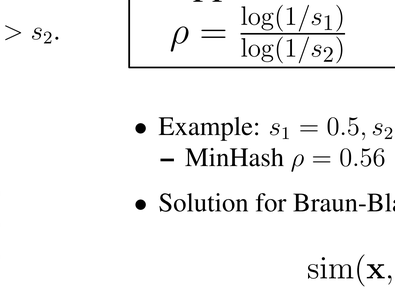


Fig: $\rho_{\text{new}} - \rho_{\text{angular}}$ for $0 < s_2 < s_1 < 1$

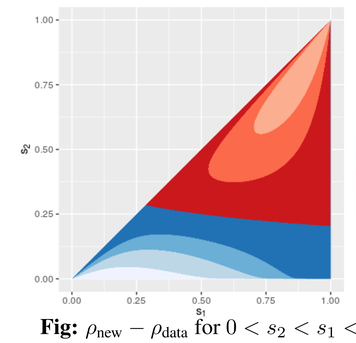


Fig: $\rho_{\text{new}} - \rho_{\text{data}}$ for $0 < s_2 < s_1 < 1$

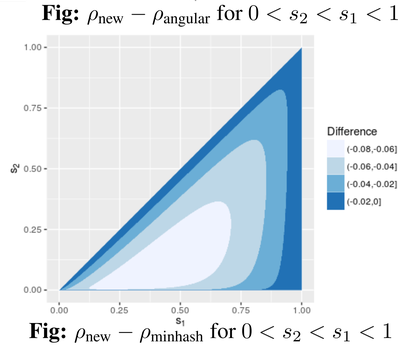


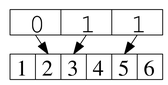
Fig: $\rho_{\text{new}} - \rho_{\text{minhash}}$ for $0 < s_2 < s_1 < 1$

Lower bound

- (r, cr)-near neighbor problem in Hamming space:
  - Preprocess a collection P of n points in $\{0, 1\}^D$.
  - Given a query point $\mathbf{q} \in \{0, 1\}^D$, if there exists $\mathbf{x} \in P$ with $\|\mathbf{q} - \mathbf{x}\|_1 \leq r$ return $\mathbf{x}' \in P$ with $\|\mathbf{q} - \mathbf{x}'\|_1 < cr$.
- [OWZ'11] lower bound: $\rho \geq 1/c$ for space $n^{1+\rho}$, time n^{ρ} solution.
- Suppose a solution for the t-regular (s_1, s_2)-similarity problem with:

$$\rho < \log(1/s_1)/\log(1/s_2)$$

- (r, cr)-nn. in Hamming space as set similarity search:
  - $\tilde{\mathbf{q}}, \tilde{\mathbf{x}} \in \{0, 1\}^D$ with $\|\tilde{\mathbf{q}} - \tilde{\mathbf{x}}\|_1 = \Delta$.
  - Define $\mathbf{x} = \{2j - \tilde{\mathbf{x}}_j \mid j \in [D]\}$.
  - Tensor: $\mathbf{x}^{\otimes\tau} = \{(j_1, \ldots, j_\tau) \mid j_i \in \mathbf{x}\}$.
  - $|\mathbf{q}^{\oplus\tau}| = |\mathbf{x}^{\oplus\tau}| = D^\tau$ and $\mathrm{sim}(\mathbf{q}^{\oplus\tau}, \mathbf{x}^{\oplus\tau}) = (1 - \Delta/D)^\tau$.



- Choose D, r, cr, τ s.t. $(1 - r/D)^\tau \approx s_1$ and $(1 - r/D)^\tau \approx s_2$.

$$\rho < \frac{\log(1/s_1)}{\log(1/s_2)} \approx \frac{\log(1/(1 - r/D))}{\log(1/(1 - cr/D))} \leq 1/c$$

Sets of different sizes

- General set similarity search:
  - Query sets have size $|\mathbf{q}| = t$.
  - Data sets have size $|\mathbf{x}| = t'$.
  - $\mathrm{sim}(\mathbf{q}, \mathbf{x}) = |\mathbf{q} \cap \mathbf{x}|/\min(|\mathbf{q}|, |\mathbf{x}|)$.

Figure: Lowest ρ-value when $|\mathbf{q}| = (1/2)|\mathbf{x}|$.