

Distance Sensitive Bloom Filters

without false negatives

Mayank Goswami[†], Rasmus Pagh[◇]
Francesco Silvestri[◇] & Johan Sivertsen[◇]

[†]City University of New York (CUNY)

[◇]IT UNIVERSITY OF COPENHAGEN

October 31, 2016

Approximate membership

Given a set S of vectors from $\{0, 1\}^d$ we want a datastructure that when queried with some $q \in \{0, 1\}^d$ will answer:

- ▶ 'yes' if $q \in S$
- ▶ 'no' if $q \notin S$ w.p. $> 1 - \epsilon$

Approximate membership

Given a set S of vectors from $\{0, 1\}^d$ we want a datastructure that when queried with some $q \in \{0, 1\}^d$ will answer:

- ▶ 'yes' if $q \in S$
- ▶ 'no' if $q \notin S$ w.p. $> 1 - \epsilon$

Solved by Bloom filters[Bloom,'70] using space $O(n \log \frac{1}{\epsilon})$.
Optimal for approximate membership testing[Carter et al.,'78].

Distance Sensitive Approximate membership

Given a set S of vectors from $\{0, 1\}^d$ an approximation factor c and a radius r we want a data structure that when queried with some $q \in \{0, 1\}^d$ it will answer:

- ▶ 'yes' if $D(q, S) \leq r$
- ▶ 'no' if $D(q, S) > cr$ w.p. $> 1 - \epsilon$

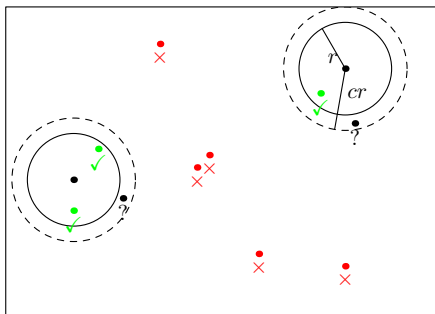


Figure: For intuition only, the box represents $\{0, 1\}^d$

Prior Work

Prior work on based on Locality Sensitive Hashing:

- ▶ Mitzenmacher & Kirsch ['06]
- ▶ Hua et. al. ['12]

Prior Work

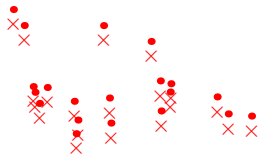
Prior work on based on Locality Sensitive Hashing:

- ▶ Mitzenmacher & Kirsch ['06]
- ▶ Hua et. al. ['12]

False negatives

Bloom filters in practice

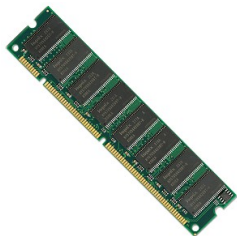
'No' → common, continue:



'Yes' → rare, check further



Query time →



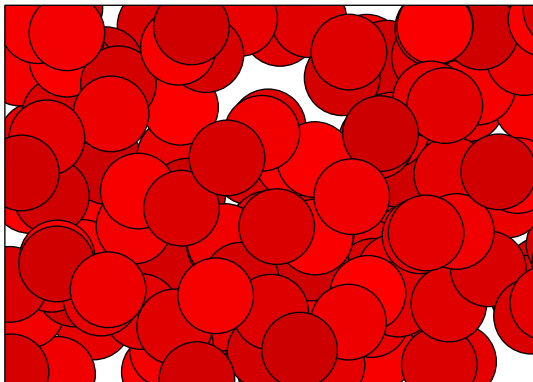
'Yes' is rare

We require $|\{UB(x, cr) : x \in S\}| < 2^{d-2}$

If

$$|\{UB(x, cr) : x \in S\}| \approx 2^d - O(n/d)$$

we only need space $O(n)$.



Theorem (General bound)

For $\epsilon < 1/4$ any distance sensitive approximate membership data structure must use space:

$$\Omega(n(r^2/d + \log 1/\epsilon))$$

Lower bound

View the data structure as a function:

$$\binom{\{0,1\}^d}{n} \rightarrow \{0,1\}^s$$

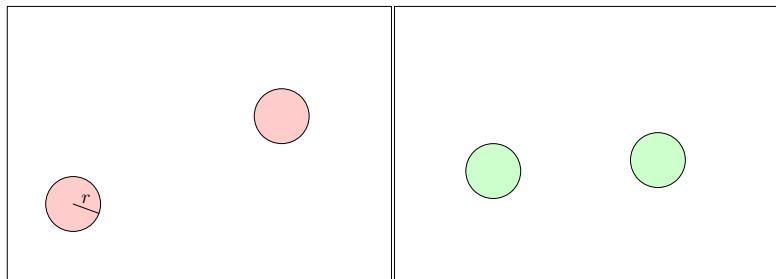


Figure: Two examples of S , $n = 2$
The boxes illustrate $\{0,1\}^d$

Lower bound

Some sets share answers.

$$\binom{\{0,1\}^d}{n} / X \rightarrow \{0,1\}^s$$

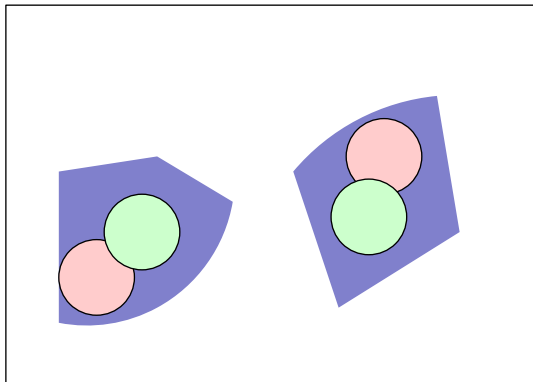
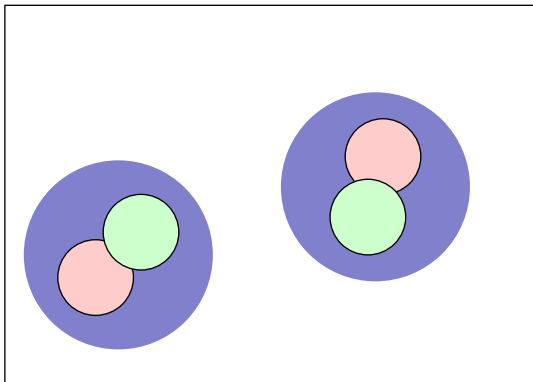


Figure: color \rightarrow 'yes', else 'no'

Lower bound

The shape covering most points is the Hamming ball (see Lemma 1).

$$\binom{\{0,1\}^d}{n} / \binom{|B^{-r}|}{n} \rightarrow \{0,1\}^s$$



Lower bound

We can bound the number of unique answers as:

$$\binom{\{0,1\}^d}{n} / \binom{|B^{-r}|}{n} \geq (\exp(2r^2/d + 1))^n$$

So

$$s \geq nr^2/d + 1$$

Lower bound

We can bound the number of unique answers as:

$$\binom{\{0,1\}^d}{n} / \binom{|B^{-r}|}{n} \geq (\exp(2r^2/d + 1))^n$$

So

$$s \geq nr^2/d + 1$$

Combine with encoding argument for $\Omega(n \log 1/\epsilon)$ based on Carter et. al. ['78] .

$$\Omega\left(n\left(r^2/d + \log \frac{1}{\epsilon}\right)\right)$$

Theorem (General bound)

For small n (fits in δcr dimensions) any distance sensitive approximate membership data structure must use space:

$$\Omega(n(r/c + \log 1/\epsilon))$$

Basic upper bound idea

Consider two vectors x, y from $\{0, 1\}^d$.

Represent the points by a signature (dot product with a random vector from $z \in \{-1, 0, 1\}^d$):

Basic upper bound idea

Consider two vectors x, y from $\{0, 1\}^d$.

Represent the points by a signature (dot product with a random vector from $z \in \{-1, 0, 1\}^d$): If $\|x - y\|_1 \leq r$:

$$z \cdot (x - y) \leq r$$

Basic upper bound idea

Consider two vectors x, y from $\{0, 1\}^d$.

Represent the points by a signature (dot product with a random vector from $z \in \{-1, 0, 1\}^d$): If $\|x - y\|_1 \leq r$:

$$z \cdot (x - y) \leq r$$

If $\|x - y\|_1 \geq cr$:

$$\Pr[|z \cdot (x - y)| > r] > f(c)$$

The signature method

Construction of the random matrix M of size $m \times d$ (m to be defined). View M as the result of d updates, each adding some value from $\{-1, 1\}$:

Example:

$$\begin{array}{cccc} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{array}$$

M , Initial state.

The signature method

Construction of the random matrix M of size $m \times d$ (m to be defined). View M as the result of d updates, each adding some value from $\{-1, 1\}$:

Example:

$$\begin{array}{cccc} \downarrow & & & \\ 0 & 0 & \dots & 0 \\ \mathbf{1} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{array}$$

M , Update 1

The signature method

Construction of the random matrix M of size $m \times d$ (m to be defined). View M as the result of d updates, each adding some value from $\{-1, 1\}$:

Example:



0	0	...	0
1	0	...	0
⋮	⋮	⋱	⋮
0	-1	...	0

Update 2

The signature method

Construction of the random matrix M of size $m \times d$ (m to be defined). View M as the result of d updates, each adding some value from $\{-1, 1\}$:

Example:

$$\begin{array}{cccc} & & & \downarrow \\ 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & \mathbf{1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & -1 & \dots & 0 \end{array}$$

M , Update d

M'

Let M' be the sub-matrix formed by the columns of M corresponding to entries where $(x - y)_i \neq 0$.

Example:

Let $(x - y) = \{0, 1, 1, 0, 0, \dots, 1\}^T$ then we get

$$M = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1d} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{md} \end{bmatrix} \rightarrow M' = \begin{bmatrix} x_{12} & x_{13} & x_{1d} \\ x_{22} & x_{23} & x_{2d} \\ \vdots & \vdots & \vdots \\ x_{m2} & x_{m3} & x_{md} \end{bmatrix}$$

M' has $D(x, y)$ non-zero entries.

Signature for $c = O(1)$

Given M and some $x \in \{0, 1\}^d$ compute the signature of x by:

$$\sigma(x)_i = (m_i \cdot x) \pmod 2 \text{ for } i \in [m]$$

Signature for $c = O(1)$

Given M and some $x \in \{0, 1\}^d$ compute the signature of x by:

$$\sigma(x)_i = (m_i \cdot x) \pmod{2} \text{ for } i \in [m]$$

Definition (Gaps)

The *gap vector* Γ is defined by:

$$\Gamma(x, y)_i = (\sigma(x)_i - \sigma(y)_i) \pmod{2} = (M(x - y))_i \pmod{2}$$

And the *gap* γ by:

$$\gamma(x, y) = \|\Gamma(x, y)\|_1$$

Central Properties of the signature

Definition (Gap)

$$\gamma(x, y)_i = \sum_{i=1}^m |M(x - y)_i \bmod 2|$$

Theorem

For each pair of vectors $x, y \in \{0, 1\}^d$:

1. if $D(x, y) \leq r$ then $\gamma(x, y) \leq r$
2. if $D(x, y) > cr$ then $\Pr[\gamma(x, y) > r] > 1 - \epsilon$.

Property 1: If $D(x, y) \leq r$ then $\gamma \leq r$

$$\gamma(x, y) = \|M(x - y) \bmod 2\|_1$$

Property 1: If $D(x, y) \leq r$ then $\gamma \leq r$

$$\gamma(x, y) = \|M(x - y) \bmod 2\|_1 = \|M'(x - y) \bmod 2\|_1$$

- ▶ Only M' contributes
- ▶ M' has $D(x, y) \leq r$ non-zero entries

Property 1: If $D(x, y) \leq r$ then $\gamma \leq r$

$$\gamma(x, y) = \|M(x - y) \bmod 2\|_1 = \|M'(x - y) \bmod 2\|_1$$

- ▶ Only M' contributes
- ▶ M' has $D(x, y) \leq r$ non-zero entries
- ▶ Each entry in $(x - y)$ "hits" one entry in M'

Property 1: If $D(x, y) \leq r$ then $\gamma \leq r$

$$\gamma(x, y) = \|M(x - y) \bmod 2\|_1 = \|M'(x - y) \bmod 2\|_1$$

- ▶ Only M' contributes
- ▶ M' has $D(x, y) \leq r$ non-zero entries
- ▶ Each entry in $(x - y)$ "hits" one entry in M'

- ▶ $M'(x - y) = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & -1 & 0 \end{bmatrix} \times \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$

Property 1: If $D(x, y) \leq r$ then $\gamma \leq r$

$$\gamma(x, y) = \|M(x - y) \bmod 2\|_1 = \|M'(x - y) \bmod 2\|_1$$

- ▶ Only M' contributes
- ▶ M' has $D(x, y) \leq r$ non-zero entries
- ▶ Each entry in $(x - y)$ "hits" one entry in M'

- ▶ $M'(x - y) = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & -1 & 0 \end{bmatrix} \times \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$

- ▶ At most the distance is preserved, it can never increase.

Property 1: If $D(x, y) \leq r$ then $\gamma \leq r$

$$\gamma(x, y) = \|M(x - y) \bmod 2\|_1 = \|M'(x - y) \bmod 2\|_1 \leq r$$

- ▶ Only M' contributes
- ▶ M' has $D(x, y) \leq r$ non-zero entries
- ▶ Each entry in $(x - y)$ "hits" one entry in M'

- ▶ $M'(x - y) = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & -1 & 0 \end{bmatrix} \times \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$

- ▶ At most the distance is preserved, it can never increase.

No false negatives!

Property 2: If $D(x, y) > cr$ then $\Pr[\gamma > r] > 1 - \epsilon$

Definition (Odd rows)

Let m'_i denote row i of M' . We call m'_i *odd* if it has an odd number of non-zero entries.

Claim 1: An odd row increases the gap:

Claim 2: There are more than r odd rows.

Property 2: If $D(x, y) > cr$ then $\Pr[\gamma > r] > 1 - \epsilon$

Definition (Odd rows)

Let m'_i denote row i of M' . We call m'_i *odd* if it has an odd number of non-zero entries.

Claim 1: An odd row increases the gap:

$$\text{If } m_i \text{ is odd, then } M(x - y)_i = \sum^{\text{odd}} \pm 1 = \text{odd}$$

so

$$|M(x - y)_i \pmod 2| = 1$$

Claim 2: There are more than r odd rows.

Property 2: If $D(x, y) > cr$ then $\Pr[\gamma > r] > 1 - \epsilon$

Claim 2: There are more than r odd rows:

Property 2: If $D(x, y) > cr$ then $\Pr[\gamma > r] > 1 - \epsilon$

Claim 2: There are more than r odd rows:

Recalling the construction of M we think of M' as the result of $D(x, y) > cr$ updates. Let Y_{cr} denote the *number of odd rows turned even* in the final cr updates. The final number of odd rows is bounded below by:

$$cr - 2Y_{cr}$$

Use a Chernoff bound to get:

$$\Pr[Y_{cr} \geq (c-1)r/2] \leq e^{-((\frac{c-1}{c})^2 \frac{m}{24} - \frac{(c-1)r}{2})}$$

where m is the number of rows in the matrix, i.e. the number of bits in each signature.

Upper bound

The storage is dominated by the number of rows in M :

$$m \geq 24 \frac{2c^2}{c-1} \max\left(r; \frac{1}{c-1} \log 1/\epsilon\right)$$

Directly determining the size of each signature.

When $c \leq 2$ this gives optimal $O(n(r/c + \log 1/\epsilon))$ bits.

Thank you!

Thank you!
Questions?

Bounding Y_{cr}

$$\Pr[Y_j = 1] \leq (Y_0 + j - 1)/m \leq 3cr/m \rightarrow$$
$$\mu = E[Y_{cr}] \leq 3(cr)^2/m$$

Bounding Y_{cr}

$$\mu = E[Y_{cr}] \leq 3(cr)^2/m$$

In the Chernoff-bound:

$$Pr[Y_{cr} \geq \mu(1 + \eta)] \leq e^{-\eta^2\mu/2}$$

Set $\eta = \frac{(c-1)r}{2\mu}$ to get:

$$Pr[Y_{cr} \geq (c-1)r/2] \leq e^{-((\frac{c-1}{c})^2 \frac{m}{24} - \frac{(c-1)r}{2})}$$

Bounding Y_{cr}

$$\mu = E[Y_{cr}] \leq 3(cr)^2/m$$

In the Chernoff-bound:

$$Pr[Y_{cr} \geq \mu(1 + \eta)] \leq e^{-\eta^2\mu/2}$$

Set $\eta = \frac{(c-1)r}{2\mu}$ to get:

$$Pr[Y_{cr} \geq (c-1)r/2] \leq e^{-((\frac{c-1}{c})^2 \frac{m}{24} - \frac{(c-1)r}{2})}$$

Now let $m \geq 24 \frac{c^2}{c-1} \max(r; \frac{2}{c-1} \log 1/\epsilon)$ to get:

$$Pr[Y_{cr} < (c-1)r/2] \geq 1 - \epsilon$$

Bounding Y_{cr}

$$\mu = E[Y_{cr}] \leq 3(cr)^2/m$$

In the Chernoff-bound:

$$Pr[Y_{cr} \geq \mu(1 + \eta)] \leq e^{-\eta^2\mu/2}$$

Set $\eta = \frac{(c-1)r}{2\mu}$ to get:

$$Pr[Y_{cr} \geq (c-1)r/2] \leq e^{-((\frac{c-1}{c})^2 \frac{m}{24} - \frac{(c-1)r}{2})}$$

Now let $m \geq 24 \frac{c^2}{c-1} \max(r; \frac{2}{c-1} \log 1/\epsilon)$ to get:

$$Pr[Y_{cr} < (c-1)r/2] \geq 1 - \epsilon$$

So the total number of odd rows is:

$$Y_0 - Y_{cr} + (cr - Y_{cr}) \geq cr - 2Y_{cr} > r$$

w.p $> 1 - \epsilon$