Distance Sensitive Bloom Filters

without false negatives

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

† max planck institut informatik

◇ IT UNIVERSITY OF COPENHAGEN

August 23, 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 $O\left(n \log \frac{1}{\epsilon}\right)$.
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. $> 1 - \epsilon$

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

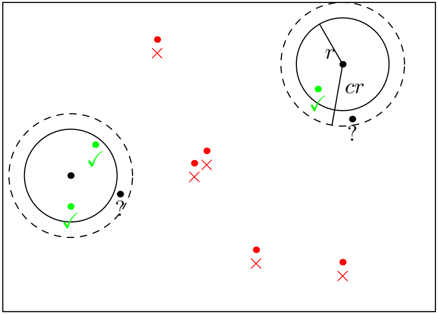


Figure: For intuition only, the box represents $\{0,1\}^d$

Prior Work

Prior work on based on Locality Sensitive Hashing:

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

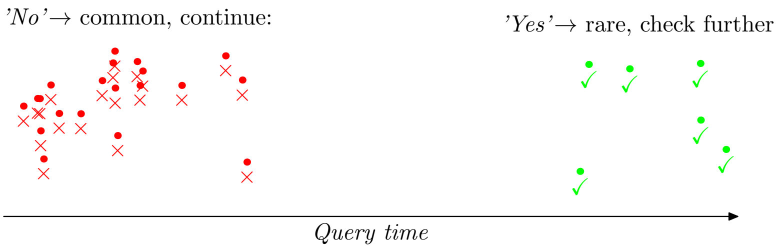
Prior Work

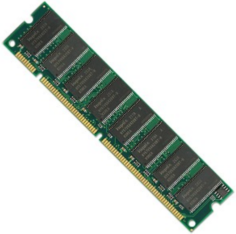
Prior work on based on Locality Sensitive Hashing:

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

False negatives

Bloom filters in practice





Overview

1. Upper bound for $c = O(1)$
2. Lower bound

Also in the paper: Average guarantees, Upper bound for all settings of c.

Basic idea

Consider two vectors x, y from $\{0,1\}^d$.
We take the dot product with a random vector from $z \in \{-1,1\}^d$:

Basic idea

Consider two vectors x, y from $\{0,1\}^d$.
We take the dot product with a random vector from $z \in \{-1,1\}^d$:

$$z \cdot x - z \cdot y = z \cdot (x - y)$$

Basic idea

Consider two vectors x, y from $\{0,1\}^d$.
We take the dot product with a random vector from $z \in \{-1,1\}^d$:

$$z \cdot x - z \cdot y = z \cdot (x - y)$$

If $\|x - y\|_1 \leq r$:

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

Basic idea

Consider two vectors x, y from $\{0,1\}^d$.
We take the dot product with a random vector from $z \in \{-1,1\}^d$:

$$z \cdot x - z \cdot y = z \cdot (x - y)$$

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:

$$\begin{matrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{matrix}$$

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:

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

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:

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

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:

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

M, Update d

Signature for $c = O(1)$

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

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

Signature for $c = O(1)$

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

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

Definition (Gaps)

The *gap vector* Γ is defined by:
$\Gamma(x,y)_i = (\sigma(x)_i - \sigma(y)_i) \mod 2 = (M(x-y))_i \mod 2$
And the *gap* γ by:
$\gamma(x,y) = \|\Gamma(x,y)\|_1$

Central Properties of the signature

Definition (Gap vector)

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

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 $\Pr[\gamma(x,y) > r] > 1-\epsilon$.*

Central Properties of the signature

Definition (Gap vector)

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

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 $\Pr[\gamma(x, y) > r] > 1 - \epsilon$.*

Time for a *small* detour.

Data-structure outline

Store the set:

$$Z = \{\sigma(x) : x \in S\}, |Z| = O(nm)$$

Query with q:

- 'yes' if $\exists z \in Z$ such that $D(\sigma(q), z) \leq r$
- 'no' otherwise

M'

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

Example:

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

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

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

M'

Let M' be the sub-matrix formed by the columns of M corresponding to entries where $(x-y)_i \neq 0$.
Example:
Let $(x-y) = \{0, 1, 1, 0, 0, \cdots, 1\}^T$ then we get

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

M' has $D(x, y)$ non-zero entries. Back to the properties

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

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

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

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

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

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

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

- 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 $\gamma \leq r$

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

- 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 $\gamma \leq r$

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

- Only M' contributes
- M' has $D(x, y) \leq r$ non-zero entries
- Each entry in $(x - y)$ "'hits"' one entry in M'
- $M'(x - y) = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & -1 & 0 \end{bmatrix} \times \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$
- "Best case" they hit in seprate rows, otherwise the sum is eaten by the log.

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

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

- Only M' contributes
- M' has $D(x, y) \leq r$ non-zero entries
- Each entry in $(x - y)$ "'hits"' one entry in M'
- $M'(x - y) = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & -1 & 0 \end{bmatrix} \times \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$
- "Best case" they hit in seprate rows, otherwise the sum is eaten by the log.

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:

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

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

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_0 denote the number of odd rows when there are cr updates left. There are two cases:
1.If $Y_0 > cr + r$:

$$Y_0 > cr + r \rightarrow Y_{D(x,y)} > r$$

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

2.If $Y_0 \leq cr + r$:

Let Y_{cr} denote the *number of odd rows turned even* in the last cr updates. So the total number of odd rows is:

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

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

2.If $Y_0 \leq cr + r$:

Let Y_{cr} denote the *number of odd rows turned even* in the last cr updates. So the total number of odd rows is:

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

Use a Chernoff bound to get:

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

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

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

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

2.If $Y_0 \leq cr + r$:

Let Y_{cr} denote the *number of odd rows turned even* in the last cr updates. So the total number of odd rows is:

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

Use a Chernoff bound to get:

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

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 $> 1 - \epsilon$.

Upper bound

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

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

Directly determining the size of each signature.
When $c \leq 2$ this gives optimal $O\left(n(r/c + \log 1/\epsilon)\right)$ bits.
(We used small n, $cr \leq d/2$, $\epsilon < 1/4$).
More parameter settings covered in the paper.

Lower bound

Theorem (General bound)

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

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

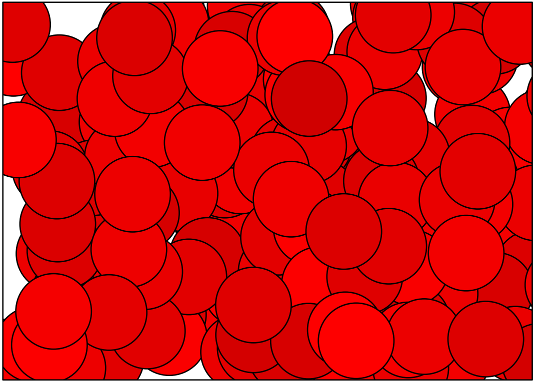
Lower bound

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

If

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

we only need space $O\left(n\right)$.



Lower bound

View the data structure as a function:

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

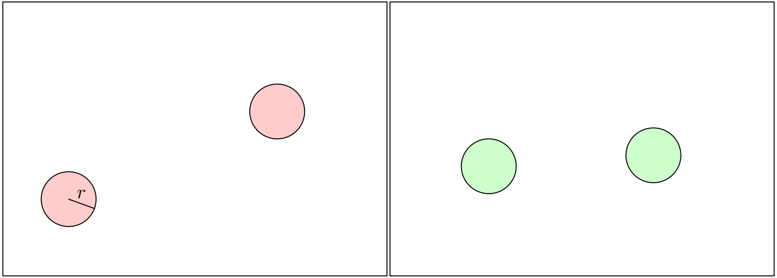


Figure: Two examples of S, $n = 2$
The boxes illustrate $\{0,1\}^d$

Lower bound

Some sets share answers.

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

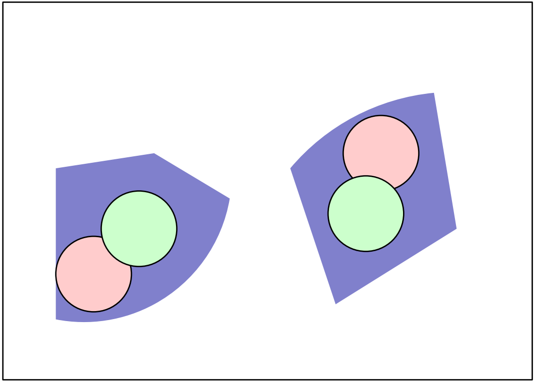
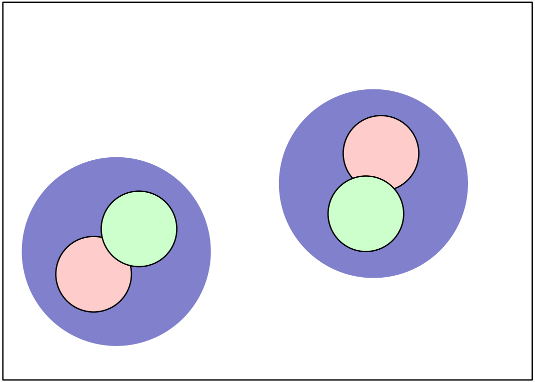


Figure: color→ 'yes', else 'no'

Lower bound

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

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



Lower bound

We can bound the number of unique answers as:

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

So

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

Lower bound

We can bound the number of unique answers as:

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

So

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

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

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

Recall upper bound

$$O\left(n(r/c + \log 1/\epsilon)\right)$$

Still a gap to:

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

Small n

Theorem (General bound)

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

$$\Omega\left(n(r/cr + \log 1/\epsilon)\right)$$

If n is small a structure for dimension d works for $d' = \delta cr < d$.

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 $> 1-\epsilon$