New Bounds for $k - hashing$

S. Costa, S. Della Fiore, M. Dalai

Università degli studi di Brescia

May 15, 2020

Overview

1. Three equivalent formulation
   - Combinatorial formulation
   - Information Theory interpretation
   - Computer Science interpretation
2. Hansel's Lemma
   - Definition
   - Application to k-hashing
3. Upper bounds on the cardinality of k-hash codes
   - Known upper bounds from literature
   - Main ideas behind Costa and Dalai's work for $k = 5, 6$
   - New bounds for $k = 6, 7, 8$

Combinatorial formulation

Problem (k-hashing)

How can we upper bound the cardinality of a set of vectors of length n over an alphabet of size k, with the property that, for every subset of k vectors there is a coordinate in which they all differ?

Combinatorial formulation

Problem (k-hashing)

How can we upper bound the cardinality of a set of vectors of length n over an alphabet of size k, with the property that, for every subset of k vectors there is a coordinate in which they all differ?

Very easy to formulate but very difficult to solve.

Information Theory and Computer Science interpretation

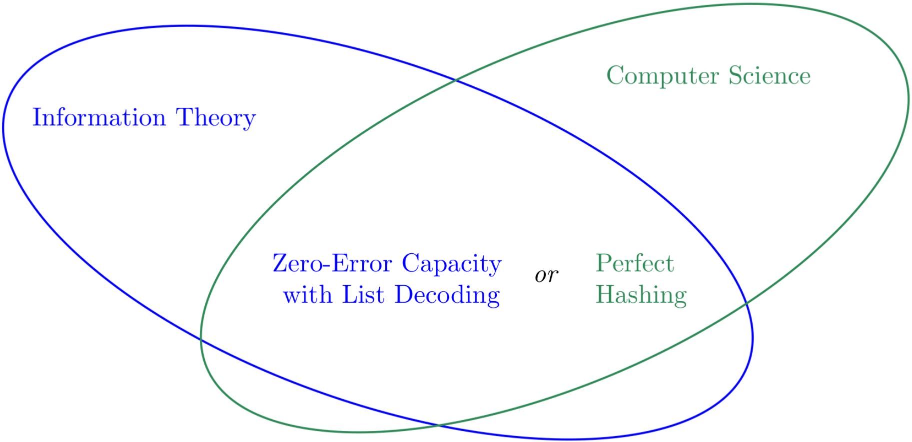


Figure: *ISIT 2017 [5].*

Zero-Error Capacity with List Decoding

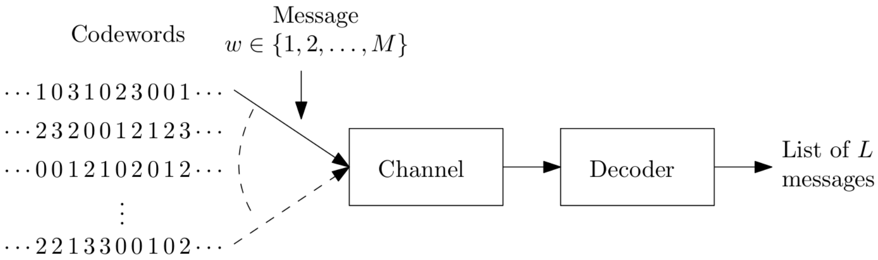


Figure: *ISIT 2017 [5].*

1. The decoder outputs a list of L messages
2. There is an **error** if the original message is not in the list
3. **Zero-error** code: the correct message is always in the list $\iff$ No $L+1$ codewords are compatible with any output sequence

Definition of Zero-Error code under List Decoding

Given a channel (bipartite-graph) $H = (V, W, E)$ where V correspond to channel inputs, W to channel outputs and $(v, w) \in E$ if w can be received when v is trasmitted.

Definition (Zero-error code under LD)

A code $C \subseteq V^n$ achieve zero-error under list-of-L decoding if for every subset $\{c^{(1)}, c^{(2)}, \ldots, c^{(L+1)}\}$ of $L+1$ codewords, there is a coordinate i such that the symbols $c_i^{(1)}, c_i^{(2)}, \ldots, c_i^{(L+1)}$ don't share a common neighbor in W.

Meaning that C is an independent set in $(L+1)$-uniform hypergraph defined on V^n where hyperedges correspond to tuples whose i'th symbols have a common neighbor in W for every i.

(see Körner-Marton 1990, "On the capacity of uniform hypergraph")

(L+1)/L Channel - Example

Let $L = 3$ then $4/3$-Channel follows:

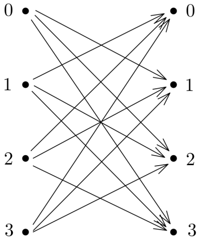


Figure: *ISIT 2017 [5].*

The four inputs have no common output meaning we can build 4-tuples which cannot be confused

$$
\begin{array}{rcccccccc}
x & = & 0 & 2 & 0 & 2 & 3 & 1 & \cdots \\
y & = & 2 & 3 & 1 & 0 & 2 & 1 & \cdots \\
z & = & 1 & 3 & 2 & 3 & 3 & 0 & \cdots \\
t & = & 1 & 0 & 3 & 2 & 1 & 2 & \cdots
\end{array}
$$

Perfect Hash function

It is an **injective** function that maps distinct elements of a set into a set of integers, with **no collision**.

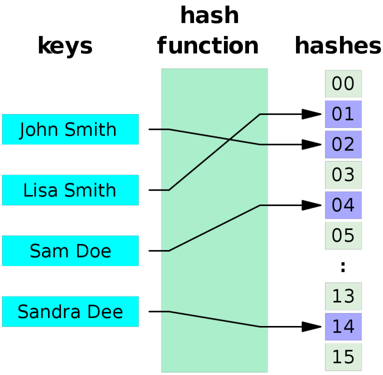


Figure: *Wikipedia.*

Perfect Hash functions for $k = 4$

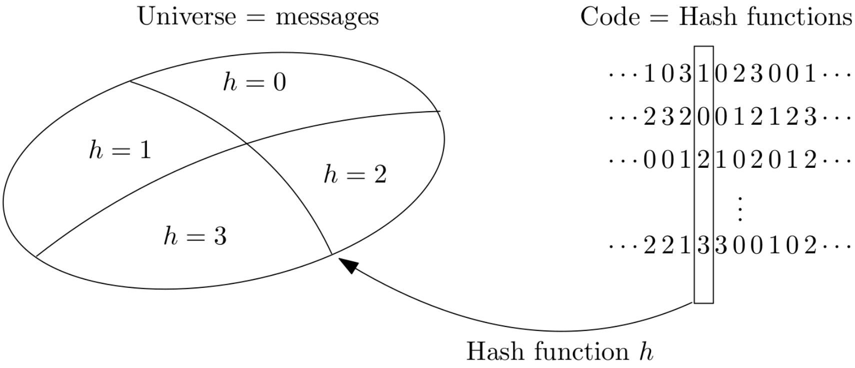


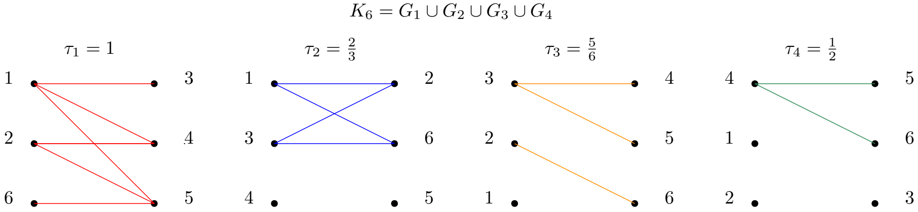
Figure: *ISIT 2017 [5].*

Perfect hashing: any x, y, z, t are separated by some hash function.

Hansel's Lemma

Let $[N] = \{1, 2, \ldots, N\}$, K_N is the complete graph on $[N]$, and

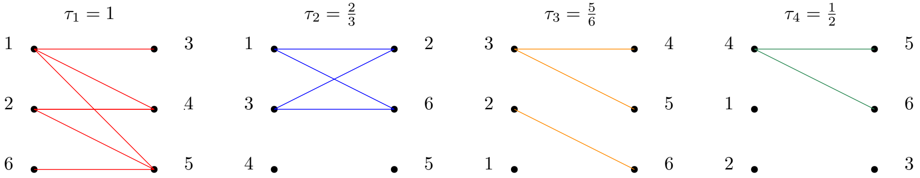
1. G_i, $i \in J$, finite sequence of bipartite graphs on $[N]$
2. τ_i is the fraction of non-isolated vertices in G_i



Let $[N] = \{1, 2, \ldots, N\}$, K_N is the complete graph on $[N]$, and

1. G_i, $i \in J$, finite sequence of bipartite graphs on $[N]$
2. τ_i is the fraction of non-isolated vertices in G_i

$$K_6 = G_1 \cup G_2 \cup G_3 \cup G_4$$



Lemma (Hansel)

If $\cup_{i \in J} G_i = K_N$ *then*

$$\log_2(N) \leq \sum_{i \in J} \tau_i$$

Hansel's Lemma for k-hashing

Let C be a set of vectors (code) that is k-separated. Given $k-2$ codewords $x_1, x_2, \dots, x_{k-2}$, let $G_i^{x_1,x_2,\dots,x_{k-2}}$ be the graph on $C \setminus \{x_1, x_2, \dots, x_{k-2}\}$ with

$$E(G_i^{x_1,x_2,\dots,x_{k-2}}) = \{(y,y') : (x_{1,i}, x_{2,i}, \dots, x_{k-2,i}, y_i, y_i') \text{ are all distinct}\}$$

1. If $|\{x_{1,i}, x_{2,i}, \dots, x_{k-2,i}\}| < k-2$ then G_i is the empty graph
2. Otherwise G_i is a **bipartite** graph

It is easy to see that $\cup_i G_i^{x_1,x_2,\dots,x_{k-2}} = K_{|C|-k+2}$ then

Hansel's Lemma for k-hashing

Let C be a set of vectors (code) that is k-separated. Given $k-2$ codewords $x_1, x_2, \ldots, x_{k-2}$, let $G_i^{x_1,x_2,\ldots,x_{k-2}}$ be the graph on $C \setminus \{x_1, x_2, \ldots, x_{k-2}\}$ with

$$E(G_i^{x_1,x_2,\ldots,x_{k-2}}) = \{(y, y') : (x_{1,i}, x_{2,i}, \ldots, x_{k-2,i}, y_i, y_i') \textit{ are all distinct}\}$$

1. If $|\{x_{1,i}, x_{2,i}, \ldots, x_{k-2,i}\}| < k-2$ then G_i is the empty graph
2. Otherwise G_i is a **bipartite** graph

It is easy to see that $\cup_i G_i^{x_1,x_2,\ldots,x_{k-2}} = K_{|C|-k+2}$ then

We can apply Hansel

Upper bound on the cardinality of k-separated sets

Given a k-separated set of vectors of length n over an alphabet of cardinality k and fixing $k-2$ vectors from C, we know thanks to Hansel's Lemma that

$$\log_2(|C|-k+2) \leq \sum_{i=1}^{n} \tau_i(x_1, x_2, \ldots, x_{k-2})$$

where $\tau_i(x_1, x_2, \ldots, x_{k-2})$ is the fraction of non-isolated vertices in $G_i^{x_1,x_2,\ldots,x_{k-2}}$.

How to get a good bound?

Upper bound on the cardinality of k-separated sets

Given a k-separated set of vectors of length n over an alphabet of cardinality k and fixing $k-2$ vectors from C, we know thanks to Hansel's Lemma that

$$\log_2(|C|-k+2) \le \sum_{i=1}^{n} \tau_i(x_1, x_2, \ldots, x_{k-2})$$

where $\tau_i(x_1, x_2, \ldots, x_{k-2})$ is the fraction of non-isolated vertices in $G_i^{x_1,x_2,\ldots,x_{k-2}}$.

How to get a good bound?

Choose $x_1, x_2, \ldots, x_{k-2}$ such that $\sum_i \tau_i$ is small.

Known upper bounds from Literature

Let $R_k = \limsup_{n\to\infty} \frac{log_2|C|}{n}$ (rate of the laregest k-hash code) then

1. Fredman-Komlós (1985) we have that $R_k \leq \frac{k!}{k^{k-1}}$ picking $x_1, x_2, \ldots, x_{k-2}$ uniformly at random from the code

Known upper bounds from Literature

Let $R_k = \limsup_{n\to\infty} \frac{log_2|C|}{n}$ (rate of the laregest k-hash code) then

1. Fredman-Komlós (1985) we have that $R_k \leq \frac{k!}{k^{k-1}}$ picking $x_1, x_2, \ldots, x_{k-2}$ uniformly at random from the code
2. Arikan (1994) for $k = 4$ picking x_1, x_2 with small Hamming distance we have that $R_4 \leq 0.3512$ (using Plotkin bound)

Known upper bounds from Literature

Let $R_k = \limsup_{n\to\infty} \frac{log_2|C|}{n}$ (rate of the laregest k-hash code) then

1. Fredman-Komlós (1985) we have that $R_k \leq \frac{k!}{k^{k-1}}$ picking $x_1, x_2, \ldots, x_{k-2}$ uniformly at random from the code
2. Arikan (1994) for $k = 4$ picking x_1, x_2 with small Hamming distance we have that $R_4 \leq 0.3512$ (using Plotkin bound)
3. Dalai, Guruswami, Radhakrishnan (2017) for $k = 4$ mixing the previous two ideas we have that $R_4 \leq 6/19 \approx 0.3158$

Known upper bounds from Literature

Let $R_k = \limsup_{n\to\infty} \frac{log_2|C|}{n}$ (rate of the laregest k-hash code) then

1. Fredman-Komlós (1985) we have that $R_k \leq \frac{k!}{k^{k-1}}$ picking $x_1, x_2, \ldots, x_{k-2}$ uniformly at random from the code
2. Arikan (1994) for $k = 4$ picking x_1, x_2 with small Hamming distance we have that $R_4 \leq 0.3512$ (using Plotkin bound)
3. Dalai, Guruswami, Radhakrishnan (2017) for $k = 4$ mixing the previous two ideas we have that $R_4 \leq 6/19 \approx 0.3158$
4. Guruswami, Riazanov (2018) improve for every k the bound of F-K. They compute the value only for $k = 5, 6$ (for $k > 6$ modulo a conjecture).

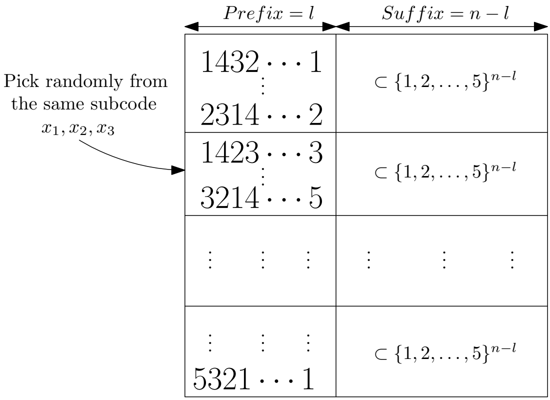
Known upper bounds from Literature

Let $R_k = \limsup_{n\to\infty} \frac{log_2|C|}{n}$ (rate of the laregest k-hash code) then

1. Fredman-Komlós (1985) we have that $R_k \leq \frac{k!}{k^{k-1}}$ picking $x_1, x_2, \ldots, x_{k-2}$ uniformly at random from the code
2. Arikan (1994) for $k = 4$ picking x_1, x_2 with small Hamming distance we have that $R_4 \leq 0.3512$ (using Plotkin bound)
3. Dalai, Guruswami, Radhakrishnan (2017) for $k = 4$ mixing the previous two ideas we have that $R_4 \leq 6/19 \approx 0.3158$
4. Guruswami, Riazanov (2018) improve for every k the bound of F-K. They compute the value only for $k = 5, 6$ (for $k > 6$ modulo a conjecture).
5. Costa, Dalai (2020) for $k = 5, 6$ we have that $R_5 \leq 0.1697$ and $R_6 \leq 0.0875$

Costa, Dalai (2020) - 1

Main idea is to construct a family of subcodes Ω such that any $k-2$ codewords of a given subcode collide in all coordinates from 1 to l.
Example for $k = 5$:



Costa, Dalai (2020) - 2

Some constraints to keep in mind:

1. If Ω is a partition of $\{1, 2, \ldots, k\}^l$ then $|\Omega| \leq \left\lfloor \left(\frac{k}{k-3}\right)^{l(1+o(1))} \right\rfloor$ if for all $w \in \Omega$ and $i = 1, 2, \ldots, l$, the i-th projection of w has cardinality at most $k-3$.
2. If $l \leq \frac{nR - 2\log_2 n}{\log(\frac{k}{k-3})}$, we can consider asymptotically only subcodes C_w such that $|C_w| \geq n$.

Costa, Dalai (2020) - strategy

First choose a subcode C_w with probability $\lambda_w = \frac{|C_w|}{|C|}$, then pick uniformly at random $x_1, x_2, \ldots, x_{k-2}$ from C_w.

$$\log_2(|C| - k + 2) \leq \mathbb{E}_{w \in \Omega}[\mathbb{E}[\sum_{i=l+1}^{n} \tau_i(x_1, x_2, \ldots, x_{k-2})]]$$
$$= \sum_{i=l+1}^{n} \mathbb{E}_{w \in \Omega}[\mathbb{E}[\tau_i(x_1, x_2, \ldots, x_{k-2})]]$$

If $x_{1,i}, x_{2,i}, \ldots, x_{k-2,i}$ are distinct then
$\tau_i(x_1, x_2, \ldots, x_{k-2}) = \frac{|C|}{|C|-k+2}(1 - \sum_{j=1}^{k-2} f_{i,x_{(j,i)}})$ where f_i is the empirical probability distribution on the i-th coordinate.
Otherwise $\tau_i(x_1, x_2, \ldots, x_{k-2}) = 0$.

The ψ function

Definition (ψ function)

Given two probability vectors $p = (p_1, p_2, \ldots, p_k)$ *and* $q = (q_1, q_2, \ldots, q_k)$

$$\psi(p, q) = \sum_{\sigma \in S_k} p_{\sigma(1)} p_{\sigma(2)} \cdots p_{\sigma(k-2)} q_{\sigma(k-1)}$$

$$\mathbb{E}[\tau_i(x_1, x_2, \ldots, x_{k-2})] = (1 + o(1))\psi(f_{i|w}, f_i)$$

At the end we get

$$\mathbb{E}_{w \in \Omega}[\mathbb{E}[\tau_i(x_1, x_2, \ldots, x_{k-2})]] = (1 + o(1)) \sum_{w \in \Omega} \lambda_w \psi(f_{i|w}, f_i)$$

A clever symmetrization - The Ψ function

Since ψ is linear its second variable, we have that

$$\mathbb{E}_{w\in\Omega}[\mathbb{E}[\tau_i(x_1,x_2,\dots,x_{k-2})]] = (1+o(1))\frac{1}{2}\sum_{w,\mu\in\Omega}\lambda_w\lambda_\mu\left(\psi(f_{i|w},f_{i|\mu})+\psi(f_{i|\mu},f_{i|w})\right)$$

Definition (Ψ function)

Given two probability vectors $p=(p_1,p_2,\dots,p_k)$ *and* $q=(q_1,q_2,\dots,q_k)$

$$\Psi(p;q)=\psi(p,q)+\psi(q,p) = \sum_{\sigma\in S_k} p_{\sigma(1)}p_{\sigma(2)}\cdots p_{\sigma(k-2)}q_{\sigma(k-1)}+q_{\sigma(1)}q_{\sigma(2)}\cdots q_{\sigma(k-2)}p_{\sigma(k-1)}$$

Upper bound for the rate of k-hash codes

Let M_k be the maximum of Ψ over probability vectors p and q, then

$$log_2(|C|) \leq (1 + o(1))\frac{1}{2}(n - l) \sum_{w,\mu \in \Omega} \lambda_w \lambda_\mu M_k$$
$$= (1 + o(1))\frac{1}{2}(n - l) M_k$$

Setting $l = \left\lfloor \frac{nR - 2\log_2 n}{\log(\frac{k}{k-3})} \right\rfloor$ we get as $n \to \infty$ that

$$R_k \leq \left(\frac{2}{M_k} + \frac{1}{\log(k/(k-3))} \right)^{-1}$$

In which point Ψ attains the maximum?

Thanks to different properties of the Ψ function, we can restrict the number of points in which Ψ attains the maximum (independently from k) and then we can test each one with Mathematica (or by hand...).

Theorem (Costa, Dalai)

$k = 5$ *the maximum is at* $(\gamma, \delta, \dots, \delta; 0, \frac{1}{4}, \dots, \frac{1}{4})$ *where* $\delta = 1/44(4 + \sqrt{5})$
$k = 6$ *the maximum is at* $(1, 0, \dots, 0; 0, \frac{1}{5}, \dots, \frac{1}{5})$

Conjecture (Costa, Dalai)

For $k > 6$ *the global maximum of the* Ψ *function is at*

$$(1, 0, \dots, 0; 0, \frac{1}{k-1}, \dots, \frac{1}{k-1})$$

How to extend the work also for $k = 7, 8$?

Introducing a parameter $0 < \epsilon < \frac{1}{k-1}$ that clusterize the probability distributions of subcodes into "balanced" and "unbalanced" categories.

We have 4 different cases of (p, q) pairs, each associated with its maximum of Ψ (dependent on ϵ):

1. **balanced-balanced** $\rightarrow M_1$
2. **unbalanced-balanced** $\rightarrow M_2$
3. **unbalanced-unbalanced on a different coordinate** $\rightarrow M_3$
4. **unbalanced-unbalanced on the same coordinate** $\rightarrow M_4$

New upper bounds for $k = 6, 7, 8$

$$\mathbb{E}_{\omega\in\Omega}[\mathbb{E}[\tau_i(x_1, x_2, \dots, x_{k-2})]]$$

$$\leq$$

$$\frac{1}{2}\left[\sum_{\omega,\mu\in\Omega_b}\lambda_\omega\lambda_\mu M_1 + 2\sum_{\omega\in\Omega_b,\mu\in\Omega_u}\lambda_\omega\lambda_\mu M_2 + \sum_{i=1}^{k}\sum_{\omega,\mu\in\Omega_i}\lambda_\omega\lambda_\mu M_3 + 2\sum_{i<j}\sum_{\omega\in\Omega_j,\mu\in\Omega_i}\lambda_\omega\lambda_\mu M_4\right]$$

$$=$$

$$\lambda_0^2 M_1 + 2\lambda_0(1-\lambda_0)M_2 + \frac{(1-\lambda_0)^2}{k}M_3 + (1-\lambda_0)^2 M_4 = f(\lambda_0)$$

$$\leq$$

$$\max_{0\leq\lambda_0\leq 1} f(\lambda_0) = M$$

where $\lambda_0 = \sum_{\omega\in\Omega_b}\lambda_\omega$.

New upper bounds for $k = 6, 7, 8$

$$\mathbb{E}_{\omega\in\Omega}[\mathbb{E}[\tau_i(x_1, x_2, \ldots, x_{k-2})]]$$

$$\leq$$

$$\frac{1}{2}\left[\sum_{\omega,\mu\in\Omega_b}\lambda_\omega\lambda_\mu M_1 + 2\sum_{\omega\in\Omega_b,\mu\in\Omega_u}\lambda_\omega\lambda_\mu M_2 + \sum_{i=1}^{k}\sum_{\omega,\mu\in\Omega_i}\lambda_\omega\lambda_\mu M_3 + 2\sum_{i<j}\sum_{\omega\in\Omega_j,\mu\in\Omega_i}\lambda_\omega\lambda_\mu M_4\right]$$

$$=$$

$$\lambda_0^2 M_1 + 2\lambda_0(1-\lambda_0)M_2 + \frac{(1-\lambda_0)^2}{k}M_3 + (1-\lambda_0)^2 M_4 = f(\lambda_0)$$

$$\leq$$

$$\max_{0\leq\lambda_0\leq 1} f(\lambda_0) = M$$

where $\lambda_0 = \sum_{\omega\in\Omega_b}\lambda_\omega$.

GOAL $\rightarrow$ get a small M changing ϵ

New upper bounds for $k = 6, 7, 8$

In this case we upper bound the quadratic form and we get the following bound that depends on M

$$R_k \leq \left(\frac{2}{M} + \frac{1}{\log(k/(k-3))} \right)^{-1}$$

Theorem (Costa, Della Fiore, Dalai)

Given a k-separated set of vectors C the rates R_k for $k = 6, 7, 8$ are upper bounded as follow

$$M \approx 0.1866 \to R_6 \leq 0.08488 \text{ vs } R_6^{FK} \leq 0.09259, R_6^{CD} \leq 0.08759$$
$$M \approx 0.0861594 \to R_7 \leq 0.040898 \text{ vs } R_7^{FK} \leq 0.04284, R_7^{G} \leq 0.04279$$
$$M \approx 0.0388599 \to R_8 \leq 0.018889 \text{ vs } R_8^{FK} \leq 0.01923, R_8^{G} \leq 0.01922$$

References

- M.L. Fredman and J. Komlós, "On the size of separating systems and families of perfect hash functions", *1984*.
- J. Korner and K. Marton, "New bounds for perfect hashing via information theory", *1988*.
- E. Arikan, "An upper bound on the zero-error list-coding capacity", *1993*.
- A. Nilli, "Perfect hashing and probability", *1994*.
- M. Dalai, V. Guruswami, and J. Radhakrishnan, "An improved bound on the zero-error list-decoding capacity of the 4/3 channel", *2017-2020*.
- S. Costa, M. Dalai, "New bounds for perfect k-hashing", *2020*.