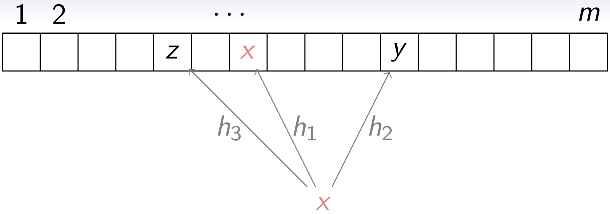
Tight Tresholds for Cuckoo Hashing via XORSAT

Martin Dietzfelbinger[1] Andreas Goerdt Michael Mitzenmacher Andrea Montanari Rasmus Pagh Michael Rink[1]

15.06.2010

[1]Research supported by DFG grant DI 412/10-1.

k-ary Cuckoo Hashing



[Fotakis, Pagh, Sanders, and Spirakis, 2005]:

- n keys, set $S \subseteq U$
- k hash functions $h_i : U \to [m] = \{1, \ldots, m\}$
- Store x in one of $T[h_i(x)]$, $i = 1, \ldots, k$.
- Maximum one key per cell.

If this is possible for all $x \in S$: Constant lookup time.
Here only: **Static case**.

Placement

For simpler notation: Key set is $S = [n]$.

For key $i \in S$ the set $A_i = \{h_1(i), h_2(i), \ldots, h_k(i)\}$ is a (fully) random subset of $[m]$ of size k.

Definition

C.H. works for $(A_i)_{1 \leq i \leq n}$
if there is a injective mapping $\sigma\colon [n] \to [m]$
such that $\sigma(i) \in A_i$, for all $i \in [n]$.
(Can store keys from S with no collision.)

Thresholds (1)

C.H. for $k = 2$: [Pagh and Rodler, 2001, 2004]

- Well understood.
- For $\frac{n}{m} < 0.5$: Pr(C.H. works) $= 1 - o(1)$.
- Related to appearance of giant connect component in cuckoo graph.

C.H. for $k \geq 3$:

Theorem ([Fotakis et al., 2005])

There are $C_1 > C_2 > 0$ such that:

$$\frac{n}{m} \leq 1 - e^{-C_2 \cdot k}, m \to \infty \Rightarrow \Pr(C.H.\ works) = 1 - o(1)\ ,$$
$$\frac{n}{m} \geq 1 - e^{-C_1 \cdot k}, m \to \infty \Rightarrow \Pr(C.H.\ works) = o(1)\ .$$

Thresholds (2)

Would like: Sharp Thresholds c_k for $k \geq 3$,
that is $c_k < 1$ such that for all c:

$$\frac{n}{m} \leq c < c_k,\ m \to \infty \quad \Rightarrow \quad \Pr(\text{C.H. works}) = 1 - o(1)\ ,$$
$$\frac{n}{m} \geq c > c_k,\ m \to \infty \quad \Rightarrow \quad \Pr(\text{C.H. works}) = o(1)\ .$$

Known in 2008:

- [Bohman and Kim, 2006]: Solution for $k = 4$.
- [Dietzfelbinger and Pagh, 2008]: Quite good lower bounds ($\approx 1 - 1.45e^{-k}$) for c_k via a result by [Calkin, 1997] on the rank of certain random matrices.

Solutions

Summer/Fall 2009:

- [Fountoulakis and Panagiotou, 2009] (arXiv, ICALP '10)
- [Frieze and Melsted, 2009] (arXiv)
- [DGMMPR 2009] (arXiv, ICALP '10)

independently solve the problem.

- [Gao and Wormald, 2010] solve a closely related problem. (No overlap in the results but in the methods.)

Outline

Next ...

A Problem with Many Faces

k-ary C.H. works

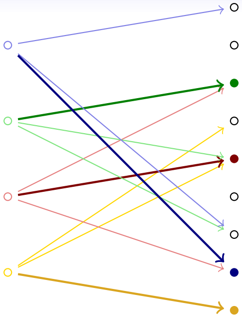
$\overset{\text{def}}{=}$ Injective mapping of keys to table cells

$\Leftrightarrow$ Left perfect matching in random bipartite graphs with left degree k

$\Leftrightarrow$ Edge orientation in random k-uniform hypergraphs

$\Leftrightarrow$ 1-submatrices in random matrices with rows of weight k

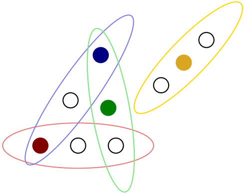
Random Bipartite graph $\mathcal{B}^k_{m,n}$



- m right nodes, n left nodes with neighbor sets of size k.
- Neighbor sets for left nodes chosen independently at random.

Question: "$\exists$ left-perfect matching for $\mathcal{B}^k_{m,n}$?"
Is there a matching in $\mathcal{B}^k_{m,n}$ that covers all left nodes?

Random Hypergraph $\mathcal{H}^k_{m,n}$



- Node set $[m]$, n hyperedges of size k.
- Hyperedges chosen independently at random.

Question: "Is $\mathcal{H}^k_{m,n}$ 1-orientable?"

Can one "direct" each hyperedge e in $\mathcal{H}^k_{m,n}$ towards one of its nodes such that each node is used for at most one edge?

Random Matrix $\mathcal{M}^k_{n,m}$

	[m]							
	1	2	3	4	5	6	7	8
A_1:	0	1	0	0	1	0	0	1
A_2:	1	0	1	0	0	0	1	0
A_3:	0	1	0	0	1	0	1	0
A_4:	0	1	1	1	0	0	0	0
A_5:	1	0	1	0	0	1	0	0

- $n \times m$ matrix $\mathcal{M}^k_{n,m}$ over $\{0,1\}$
- Rows of weight (number of 1's) exactly k, chosen randomly.

Question: "$\exists$ submatrix $\geq$ permutation matrix?"
Is there an injective mapping $\sigma\colon [n] \to [m]$ such that $\mathcal{M}^k_{n,m}(i, \sigma(i)) = 1$ for all i?

Relationship

Obvious:

- k-ary C.H.,
- degree-k left-perfect matching in bipartite graphs,
- k-uniform hypergraph orientation,
- weight-k-rows permutation submatrix

are just reformulations of the same problem.
They have the same threshold density (if any).

Next ...

The 2-core

Algorithm 1: Peeling Hypergraph

Input: $\mathcal{H}^k_{m,n} = (V, E)$

while $\exists\ v \in V$ that is covered by exactly one edge $e \in E$ **do**

  direct e towards v; `// log information`

  delete e and v;

Output: Maximal subhypergraph $\mathcal{C}^k_{\hat{m},\hat{n}}$ with min-degree ≥ 2: The "2-core".

Analogous procedure in other formulations: Always get the (equivalent) "2-core".

Relationship

k-ary C.H. works:

$\Leftrightarrow$ edges of 2-core of corr. hypergraph can be 1-oriented

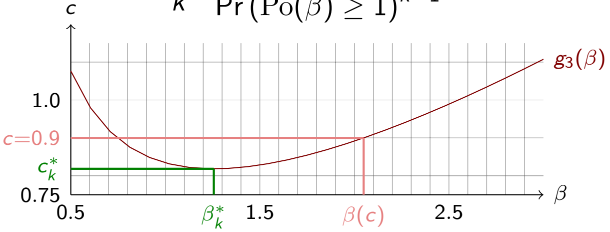
$\Leftrightarrow$ 2-core of corr. bipartite graph has a left-perfect matching

$\Leftrightarrow$ 2-core of corr. matrix has a injective mapping σ : rows $\rightarrow$ cols with entry $(i, \sigma(i)) = 1$ for all rows i in the 2-core.

Appearance 2-core

Analysis of 2-core by
[Molloy, 2005],[Cooper, 2004],[Dubois and Mandler, 2002] and others.

$$g_k(\beta) := \frac{1}{k} \cdot \frac{\beta}{\Pr\left(\mathrm{Po}(\beta) \geq 1\right)^{k-1}}, \ \beta > 0 \ .$$



- Convex in $(0, \infty) \Rightarrow \exists$ local=global minimum (β_k^*, c_k^*).
- For $c < c_k^*$ the 2-core of $\mathcal{H}_{m,n}^k$ is empty w.h.p..
- For $c > c_k^*$ there is unique $\beta(c)$ right of the β_k^* s.t. $g_k(\beta(c)) = c$.

Edge Density of 2-core

Theorem ([Molloy, 2005],[Cooper, 2004],[Dubois and Mandler, 2002],. . .)

Given $c = \frac{n}{m}$ of $\mathcal{H}^k_{m,n}$, then the edge density $\frac{\hat{n}}{\hat{m}}$ of the 2-core $\mathcal{C}^k_{\hat{m},\hat{n}}$ is tightly concentrated around

$$f(\beta(c)) = \frac{\beta(c) \cdot \Pr\left(\mathrm{Po}(\beta(c)) \geq 1\right)}{k \cdot \Pr\left(\mathrm{Po}(\beta(c)) \geq 2\right)}.$$

Definition

Let c_k be the unique c for which it holds: $f(\beta(c)) = 1$.

"c_k is the density $\frac{n}{m}$ of the hypergraph, where the density $\frac{\hat{n}}{\hat{m}}$ of the 2-core is 1."

2-core Density and Cuckoo Hashing

Let $\frac{n}{m} > c_k$:

$\Rightarrow$ Edge density in the 2-core of $\mathcal{H}^k_{m,n}$ is $\geq 1 + \delta(c)$.

$\Rightarrow$ C. H. can't work! (more egdes/keys than nodes/buckets)

Let $\frac{n}{m} < c_k$:

$\Rightarrow$ Edge density in the 2-core of $\mathcal{H}^k_{m,n}$ is $\leq 1 - \delta(c)$.

$\Rightarrow$? (Need edge density ≤ 1 for all subhypergraphs of the 2-core.)

Next . . .

1. Equivalent Formulations
2. Role of 2-Cores
3. **Thresholds for Cuckoo Hashing**
4. Extensions

Interesting Case: $\frac{\hat{n}}{\hat{m}} \leq 1$

Other works - the stony path

$\left\{ \begin{array}{l} \text{[Fountoulakis/Panagiotou 2009]} \\ \text{[Frieze/Melsted 2009]} \end{array} \right\}$ show by direct calculations

that if $\left\{ \begin{array}{l} \frac{\hat{n}}{\hat{m}} \leq 1 - \delta \\ \hat{m} = \hat{n} \end{array} \right\}$ then w.h.p. there is no subhypergraph with

edge density > 1. $\Rightarrow$ C. H. works! $\left(\left\{ \begin{array}{c} 12 \\ 20 \end{array} \right\} \text{ pages of calculations.} \right)$

Our choice - the lazy way

We show that

- the (essentially) known density thresholds for Random k-XORSAT are the same as for k-ary cuckoo hashing
- the thresholds are at the place where the edge density of the 2-core of the relevant hypergraph grows beyond 1.

Random k-XORSAT

$$(\overline{X}_1 \oplus X_2 \oplus \overline{X}_4) \wedge (\overline{X}_2 \oplus \overline{X}_4 \oplus X_5) \wedge (X_3 \oplus \overline{X}_4 \oplus X_5)$$

- n clauses, m variables
- k literals per clause

Question: "$\exists$ an assignment $x = (x_1, \dots, x_m)$ that gives all clauses value 1?"

Equivalent: Solvability of random sparse system $\mathcal{M}^k_{n,m} \cdot x = b$.
(Note: $\overline{X} = 1 \oplus X$)

$$X_1 \oplus X_2 \oplus X_4 = 1 \text{ and } X_2 \oplus X_4 \oplus X_5 = 1 \text{ and } X_3 \oplus X_4 \oplus X_5 = 0$$

Linear System

$\mathcal{M}^k_{n,m} \cdot x = b$:

- $\mathcal{M}^k_{n,m}$ is an $n \times m$ matrix, 0-1-valued, exactly k 1's per row.
- $b \in \{0,1\}^n$ is random.

Known in k-XORSAT research (e.g.[Dubois and Mandler, 2002]): $\mathcal{M}^k_{n,m}$ is equivalent to a random hypergraph $\mathcal{H}^k_{m,n}$.

Peeling off columns with exactly one 1 and the corresponding rows

- Does not change the solvability of the system.
- It remains the $\hat{n} \times \hat{m}$ matrix $\mathcal{M}^k_{\hat{n},\hat{m}}$ that corresponds to the 2-core of $\mathcal{H}^k_{m,n}$ and a reduced right hand side $\hat{b}$.

The Key Step

Theorem ([Dubois and Mandler, 2002])

$$\frac{\hat{n}}{\hat{m}} \leq c < 1 \Rightarrow \Pr(\mathcal{M}^k_{\hat{n},\hat{m}} \cdot x = \hat{b} \textit{ solvable}) = 1 - o(1) \ .$$

(Claimed for all $k \geq 3$, proved for $k = 3$.)

With $\Pr\left(\mathcal{M}^k_{\hat{n},\hat{m}} \cdot x = \hat{b} \text{ solvable} \mid \mathrm{Rank}(\mathcal{M}^k_{\hat{n},\hat{m}}) < \hat{n}\right) \leq 0.5$
it follows:

$$\Pr\left(\mathcal{M}^k_{\hat{n},\hat{m}} \text{ has full row rank}\right) = 1 - o(1)$$

$$\Rightarrow \Pr\left(\mathcal{M}^k_{\hat{n},\hat{m}} \text{ has full rank } \hat{n} \times \hat{n} \text{ sub-/permutation matrix}\right) = 1 - o(1)$$

$$\Rightarrow \Pr\left(\text{cuckoo hashing works w.r.t. rows of } \mathcal{M}^k_{\hat{n},\hat{m}}\right) = 1 - o(1) \ .$$

... which is out theorem. □

Next ...

Fractional Left Degrees

Generalization of the formulas to compute threshold c_k for arbitrary degree distributions.

Question: "What is the optimal distribution?"

Theorem

Let κ_x be the expected number of hash values for key x. Then Pr(*C.H works*) *is maximized if*
κ_x is concentrated on $\{\lfloor \kappa_x \rfloor, \lfloor \kappa_x \rfloor + 1\}$.

Larger Buckets

Assume a bucket can hold up to ℓ keys instead of just 1.

Conjecture

The threshold for this to work is where the "$(\ell+1)$-core" of $\mathcal{H}^k_{m,n}$ exceeds density ℓ.

- Known for $k = 2$ and $\ell \geq 2$ [Cain et al., 2007], [Fernholz and Ramachandran, 2007].
- Recently learned: Proved for all $k \geq 3$ and ℓ sufficiently large by [Gao and Wormald, 2010].

A Linear Time Algorithm (1)

Adaption of "selfless-algorithm" [Sanders, 2004].

Algorithm 2: (k, ℓ)-Generalized Selfless

Input: Hypergraph $\mathcal{H}^k_{m,n} = (V, E)$ with m nodes and n edges.

for $t \leftarrow 1$ **to** n **do**

- $V_0 \leftarrow \{v \in V : v \text{ is incident to undirected edge}\}$;
- $E_0 \leftarrow \{e \in E : e \text{ is undirected}\}$;
- find $v \in V_0$ with smallest priority $\pi(v)$;
- **if** $\pi(v) > \ell$ **then return failure**;
- choose $e \in E_0 \cap \{e : v \in e\}$ with minimum weight $\omega(e)$;
- direct e towards v;

A Linear Time Algorithm (2)

- $\mathcal{D}(v)$ set of hyperedges directed towards node v
- $\mathcal{U}(v)$ set of undirected hyperedges incident to node v

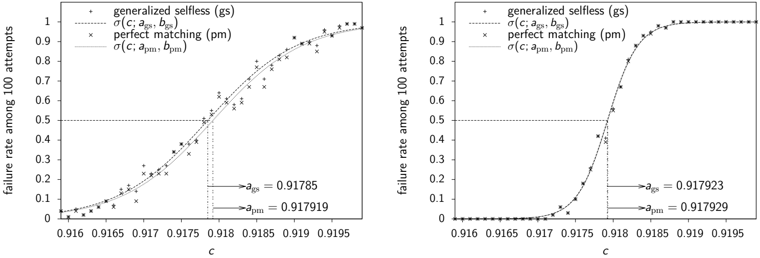
Edge weight:

$$\omega(e) \leftarrow |\{v \in e : |\mathcal{D}(v)| < \ell\}|$$

Node priority:

$$\pi(v) = \begin{cases} 0, & \text{if } |\mathcal{U}(v)| + |\mathcal{D}(v)| \leq \ell \\ \sum_{e \in \mathcal{U}(v)} \frac{1}{\omega(e)} + |\mathcal{D}(v)|, & \text{otherwise} \end{cases}$$

Generalized Selfless vs Perfect Matching



(a) $m = 10^5$ (b) $m = 10^6$

Figure : $k = 3$; theoretical threshold $c_k \approx 0.91794$, interval size 0.004

Open: Analyze one of the known dynamic versions (insertions).

Thank you!

Larger Buckets

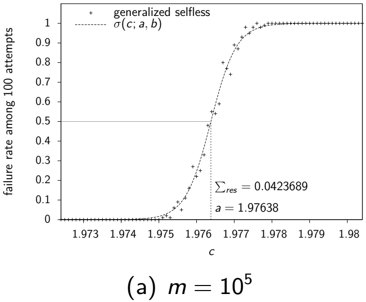


Figure : $k = 3$, $\ell = 2$; **conjectured** threshold value $c_{k,2} \approx 1.97640$

Non-integer Choices

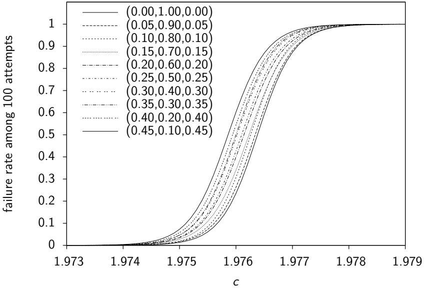


Figure : Various distributions with mean $\kappa_x = 3$; (x, y, z) stands for fraction of keys with $(k = 2, k = 3, k = 4)$; $\ell = 2, m = 10^5$.

Integral

$\ell \backslash k$	2	3	4	5	6
2	–	0.9179352767	0.9767701649	0.9924383913	0.9973795528
3	1.7940237365	1.9764028279	1.9964829679	1.9994487201	1.9999137473
4	2.8774628058	2.9918572178	2.9993854302	2.9999554360	2.9999969384
5	3.9214790971	3.9970126256	3.9998882644	3.9999962949	3.9999998884
6	4.9477568093	4.9988732941	4.9999793407	4.9999996871	4.9999999959
7	5.9644362395	5.9995688805	5.9999961417	5.9999999733	5.9999999998

Non-Integral

κ^*	$c_{\kappa^*,2}$
2.25	0.6666666667
2.50	0.8103423635
2.75	0.8788457372
3.00	0.9179352767
3.25	0.9408047937
3.50	0.9570796377
3.75	0.9685811888
4.00	0.9767701649

κ^*	$c_{\kappa^*,2}$
4.25	0.9825693463
4.50	0.9868637629
4.75	0.9900548807
5.00	0.9924383913
5.25	0.9942189481
5.50	0.9955692011
5.75	0.9965961383
6.00	0.9973795528

Bibliography (1)

Bohman, T. and Kim, J. H. (2006).
A Phase Transition for Avoiding a Giant Component.
In *Random Struct. Algorithms* 28(2), pages 195–214.

Cain, J. A., Sanders, P., and Wormald, N. C. (2007).
The Random Graph Threshold for k-orientiability and a Fast Algorithm for Optimal Multiple-Choice Allocation.
In Proc. 18th *SODA*, pages 469–476.

Calkin, N. J. (1997).
Dependent Sets of Constant Weight Binary Vectors.
In *Combinatorics, Probability & Computing* 6(3), pages 263–271.

Cooper, C. (2004).
The Cores of Random Hypergraphs with a Given Degree Sequence.
In *Random Struct. Algorithms* 25(4), pages 353–375.

Bibliography (2)

Creignou, N. and Daudé, H. (2003).
Smooth and sharp thresholds for random k-xor-cnf satisfiability.
In *ITA* 37(2), pages 127–147.

Dietzfelbinger, M. and Pagh, R. (2008).
Succinct Data Structures for Retrieval and Approximate Membership (Extended Abstract).
In Proc. 35th *ICALP* (1), pages 385–396.

Dubois, O. and Mandler, J. (2002).
The 3-XORSAT Threshold.
In Proc. 43rd *FOCS*, pages 769–778.

Fernholz, D. and Ramachandran, V. (2007).
The k-orientability Thresholds for $G_{n,p}$.
In Proc. 18th *SODA*, pages 459–468.

Bibliography (3)

Fotakis, D., Pagh, R., Sanders, P., and Spirakis, P. G. (2005).
Space Efficient Hash Tables with Worst Case Constant Access Time.
In *Theory Comput. Syst.* 38(2), pages 229–248.

Fountoulakis, N. and Panagiotou, K. (2009).
Sharp Load Thresholds for Cuckoo Hashing.
In *CoRR*, abs/0910.5147.

Frieze, A. M. and Melsted, P. (2009).
Maximum Matchings in Random Bipartite Graphs and the Space Utilization of Cuckoo Hashtables.
In *CoRR*, abs/0910.5535.

Bibliography (4)

Gao, P. and Wormald, N. C. (2010).
Load Balancing and Orientability Thresholds for Random Hypergraphs.
In Proc. 42nd *STOC*, pages 97–104.

Molloy, M. (2005).
Cores in Random Hypergraphs and Boolean Formulas.
In *Random Struct. Algorithms* 27(1), pages 124–135.

Pagh, R. and Rodler, F. F. (2001).
Cuckoo Hashing.
In Proc. 9th *ESA*, pages 121–133.

Bibliography (5)

Pagh, R. and Rodler, F. F. (2004).
Cuckoo hashing.
In *J. Algorithms* 51(2), pages 122–144.

Sanders, P.
Algorithms for Scalable Storage Servers.
In Proc. 30th *SOFSEM*, pages 82–101.