ANALYSIS OF LINEAR PROBING HASHING WITH BUCKETS (SURVEY)

ALFREDO VIOLA

(viola@fing.edu.uy)

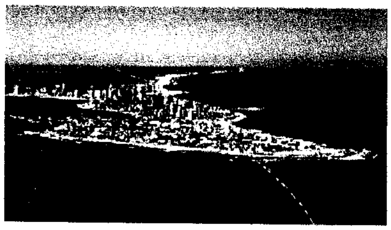
— Universidad de la República, Montevideo, URUGUAY

— LIPN - Université de Paris-Nord, Villetaneuse, FRANCE

Announcement and Call for Papers

Information Theory Workshop (ITW'2006)

March 13–17, 2006, Punta del Este, Uruguay



General Co-chairs

Gadiel Seroussi
Alfredo Viola

Program Committee

Ron Roth, co-chair
Marcelo Weinberger, co-chair
Mario Blaum
Michelle Effros
Ioannis Kontoyiannis
Hugo Krawczyk
Gábor Lugosi
Muriel Médard
Neri Merhav
Alon Orlitsky
Daniel Panario
Serap Savari
Amin Shokrollahi
Wojciech Szpankowski
Ruediger Urbanke
Alexander Vardy
Sergio Verdú

Organizing Committee

Pablo Belzarena
Gustavo Brown
Álvaro Martín
Ignacio Ramírez

Important Dates

Submission:	Oct. 31, 2005
Notification:	Dec. 23, 2005
Final version:	Jan. 15, 2006

The 2006 IEEE Information Theory Workshop (ITW 2006), organ[illegible] de Ingeniería—Universidad de la República, Uruguay and [illegible] Theory Society, will take place from the evening of Mo[illegible] Friday, March 17, 2006, at the Conrad Resort and Casino [illegible] Uruguay.

The sessions of the workshop will cover the following topics:

- Algebraic and combinatorial coding theory
- Algorithms in finite fields
- Analysis of algorithms in information theory
- Application of coding theory in computer science
- Coding techniques for storage
- Communication complexity
- Cryptography and data security
- Data compression algorithms and source coding
- Data networks
- Detection
- Information theory and statistics
- LDPC codes, turbo codes, and iterative decoding
- Multi-user information theory
- Pattern recognition and learning
- Shannon theory
- Universal prediction and on-line algorithms

Sessions will consist mainly of invited talks, but will also include slots for contributed papers. Submission of papers on the above topics is hereby solicited. The deadline for submission is **October 31, 2005**. Authors will be notified of acceptance decisions by **December 23, 2005**. The final versions of all contributions, to be published in the workshop proceedings CD, will be due by **January 15, 2006**.

Further information, such as submission guidelines, contacts, and local information, will be available at the workshop web site,

`http://www.fing.edu.uy/itw06.`

Location: Punta del Este is one of the premier resort cities in South America. It offers cultural attractions, sophistication, excellent restaurants, natural beauty, and fabulous beaches (the workshop takes place at the end of summer). It is about 1.5 hrs of scenic drive from the capital Montevideo, or a 30 min flight from Buenos Aires, Argentina.

Announcement and Call for Papers

Latin American Theoretical INformatics (LATIN'2006)

March 20–24, 2006, Valdivia, Chile

Program Committee

A. Arratia Quesada
M. Arenas
J. Correa (*Co-organizer*)
L. Devroye
G. Durán
A. Fernández
D. Fernández-Baca
A. Ferreira
F. Fomin
J. von zur Gathen
C. Gomes Fernandes
M. Golin
A. Hevia (*Co-organizer*)
M. Kiwi (*Chair*)
H. Krawczyk
E. Laber
M. Matmala
J. Matoušek
G. Navarro
D. Panario
A. Schulz
J. Sakarovitch
A. Shokrollahi
M. Singh
H. Straubing
S.-H. Teng
L. Trevisan
J. Urrutia
U. Vazirani
S. Vempala
A. Viola
Y. Wakabayashi

Important Dates

Submission:	**Sep. 21, 2005**
Notification:	**Nov. 21, 2005**
Final version:	**Dec. 16, 2005**

The 7th Latin American Theoretical INformatics Symposium will take place in March, from Monday 20 through Friday 24, 2006, in Valdivia, Chile. Submissiones in all areas of theoretical computer science are welcome, including (but not limited to):

- Algorithms
- Automata theory and formal languages
- Coding theory and data compression
- Combinatorics and graph theory
- Complexity theory
- Computational algebra
- Computational biology
- Computational geometry
- Computational number theory
- Cryptography
- Databases and information retrieval
- Data structures
- Internet and the web
- Logic in computer science
- Machine learning
- Mathematical programming
- Parallel and distributed computing
- Pattern matching
- Quantum computing
- Random structures and algorithms
- Scientific computing

Sessions will consist of keynote addresses and contributed papers. Submission of papers on the above topics is hereby solicited. The final versions of all accepted contributions will be published in the symposium proceedings by Springer-Verlag, in the Lecture Notes in Computer Science Series.

Keynote addresses will be given by:

- Ricardo **Baeza-Yates**, U. Chile
- Anne **Condon**, U. British Columbia
- Ferran **Hurtado**, U. Politècnica de Catalunya
- R. **Ravi**, Carnegie Mellon U.
- Madhu **Sudan**, MIT
- Sergio **Verdú**, Princeton U.
- Avi **Wigderson**, Institute for Advanced Study

Further information, such as submission guidelines, contacts, and local information, will be available at the conference web site,

`http://www.latin06.org.`

Location: Valdivia is located at the confluence of three rivers, 15 km (10 mi) from the coast and 840 km (525 mi) south of Santiago. It is a university town of 140,000 inhabitants, also called the "City of rivers" and "Southern Pearl." Recognized by its uniqueness due to its architectural and natural beauty, it harmonically combines Spanish and German heritage. The city proudly exhibits historical fortifications, a botanical garden, and reputed beer industry in a gorgeous natural setting.

DISTRIBUTIONAL ANALYSIS OF ROBIN HOOD LINEAR PROBING HASHING WITH BUCKETS

ALFREDO VIOLA

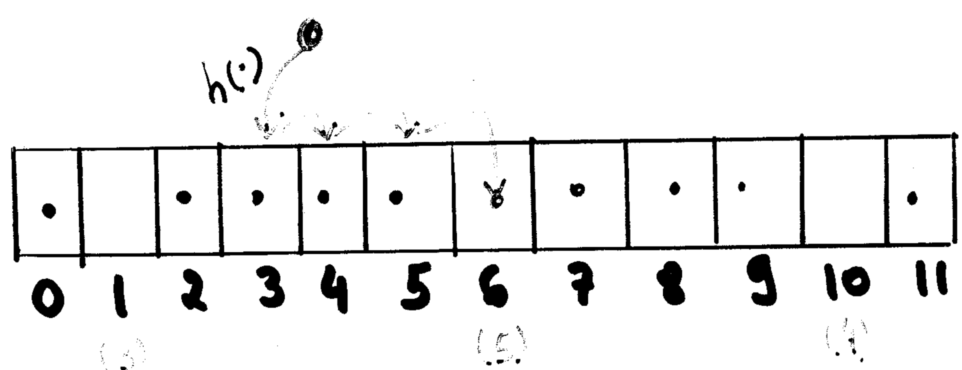
(viola@fing.edu.uy)

- Universidad de la República, Montevideo, URUGUAY.

- LIPN - Université de Paris-Nord, Villetaneuse, FRANCE.

MOTIVATION

- Linear Probing is the simplest collision resolution scheme for open addressing.
- Works well when the table is not full.
- Performance deteriorates when load factor increase.
- First proposed by Peterson in 1957.
- The analyses of problems related with linear probing show relations with other problems like tree inversions, tree paths lengths, graph connectivity, area under excursions, etc.



[illegible]

- Individual Displacements:
  - (Knuth 1962-1963). First nontrivial algorithm he analyzed.

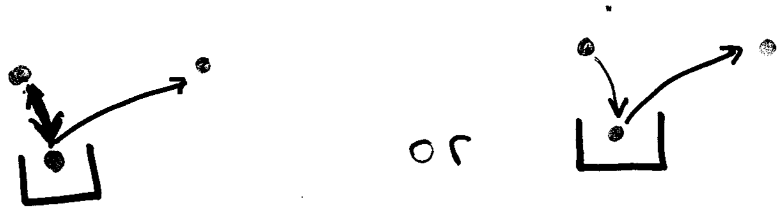
    (Konheim-Weiss 1966, Podderjugin 1967) First published analyses.
  - (Janson, Viola. To appear) Distributional Analysis for LCFS, FCFS, RH

    2005-2006
  - (Pittel 1987, Devroye-Morin-Viola) Longest probe
- Construction Cost:
  - (Knuth, Flajolet-Poblete-Viola 1998) Distributional Analysis, and relation with graph connectivity, tree inversions, tree path lenghts, area under excursions, etc. ("Airy phenomena")
  - (Janson 2001) More limit distributions.
  - (Chassaing-Louchard 2002) Phase Transitions.
- Buckets with capacity b≥1. Individual Displacements
  - (Blake-Konheim 1977) Exact PGF for FCFS and asymptotic analysis of the expected value.
  - (Poblete-Viola 1998) Exact expected value for full tables.
  - (Knuth-volume 3) expected value for α-full tables
  - (Viola 2005) Distributional Analysis of Robin Hood

    (NEW)

PRIMAL OBSERVATION WITH IMPORTANT CONSEQUENCES

- When two keys collide, either key can get the location. This can be done without looking ahead.



- Several policies can be considered without looking ahead
  - FCFS standard
  - LCFS (Poblete - Munro 1989)
  - Robin Hood ((Celis - Larson - Munro 1985) (Carlsson - Munro - Poblete 1987) → for Linear Probing.

- Neither of these heuristics change the expected value.

- Robin Hood minimizes the variance among all heuristics for Linear Probing that do not look ahead.

- $C_{b,m,n}$ → RV for number of probes in a successful search of a hash table with $\underline{m}$ buckets of capacity $\underline{b}$ and $\underline{n+1}$ elements.

- $C_{b,\alpha} = \lim C_{b,m,n}$

KNOWN RESULTS FOR ROBIN HOOD

- $C_{m,n,b} \longrightarrow$ RV for cost of successful search of a random element when we insert $\underline{n+1}$ elements in $\underline{m}$ buckets of capacity $\underline{b}$.

(Knuth 1962, Konheim-Weiss 1966)

$\left[Q_r(m,n) = \sum_{i \geq 0} \binom{i+r}{i} \frac{n^{\underline{i}}}{m^i} \right.$

- $E[C_{m,n,1}] = \frac{1}{2}\left(1 + Q_0(m,n)\right)$
- $E[C_{m,\alpha m,1}] = \frac{1}{2}\left(1 + \frac{1}{1-\alpha}\right) - \frac{1}{2(1-\alpha)^3 m} + O\left(\frac{1}{m^2}\right)$
- $E[C_{m,m,1}] = \frac{\sqrt{2\pi m}}{4} + \frac{1}{3} + \frac{1}{48}\sqrt{\frac{2\pi}{m}} + O\left(\frac{1}{m}\right)$

(Carlsson-Munro-Poblete 1987)

- $Var[C_{m,n,1}] = \frac{1}{2} Q_1(m,n) - \frac{1}{4} Q_0(m,n)^2 - \frac{1}{6} Q_0(m,n) + \frac{n}{6m} - \frac{1}{12}$
- $Var[C_{m,\alpha m,1}] = \frac{1}{4(1-\alpha)^2} - \frac{1}{6(1-\alpha)} - \frac{1}{12} + \frac{\alpha}{6} - \frac{1}{6m} - \frac{1+2\alpha}{3(1-\alpha)^4 m} + O\left(\frac{1}{m^2}\right)$
- $Var[C_{m,m,1}] = \frac{4-\pi}{8} m + \frac{1}{9} - \frac{\pi}{48} + \frac{1}{135}\sqrt{\frac{2\pi}{m}} + O\left(\frac{1}{m^2}\right) \rightarrow$ OPTIMAL!

(Janson, Viola To Appear) 2005-2006

- $C_{\alpha,1}(z) = z \; \frac{1-\alpha}{\alpha} \; \frac{e^{z\alpha} - e^{\alpha}}{z e^{\alpha} - e^{z\alpha}}$
- $C_{m,n,1}(z) = \sum_{r=1}^{n+1} \left(\sum_{i=r}^{n+1} \binom{n+1}{i} \frac{(m-n-1+i)}{(n+1)\, m^i} \sum_{k=0}^{i+r} (-1)^{i+1-r-k} \binom{i+1}{r+k} k^i \right) z^r$
- $Pr\left(\frac{C_{m,m-1,1}}{\sqrt{m}} \leq x\right) \longrightarrow$ Rayligh $(1/2)$

(Knuth Vol 3)

- $E[C_{m,\alpha m,b}] = 1 + \sum_{k \geq 1} e^{-bk\alpha} \sum_{j \geq 1} j \frac{(bk\alpha)^{bk+j-1}}{(bk+j)!}$

(Poblete - Viola 1998)

- $E[C_{m,bm-1,b}] = \frac{1}{b} \sum_{i \geq 2} \binom{bm-1}{i} \frac{(-1)^i}{m^i} \sum_{k=1}^{m} k^{i-1} \binom{bk-i}{bk-1} + \frac{m-1}{2bm} + 1$

$\ldots + O\left(\frac{1}{b^2 m}\right)$

Robin Hood - Example

- Insert records with keys 36, 77, 24, 79, 56, 69, 49, 18, 38, 97, 78, 10, 40, 70 in a table with 10 buckets of size 2 and $h(x) = x \bmod 10$

58

69	10	70		24		36	77	18	78
79	40					56	97	38	49
0	1	2	3	4	5	6	7	8	9

- Insert 58.

29

49	79	40		24		36	77	18	58
69	10	70				56	97	38	78
0	1	2	3	4	5	6	7	8	9

- Insert 29.

29	69	10	70	24		36	77	18	58
49	79	40				56	97	38	78
0	1	2	3	4	5	6	7	8	9

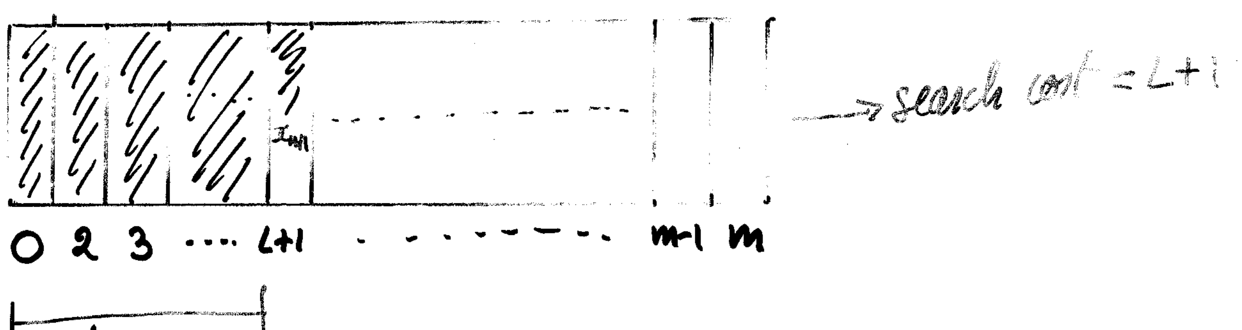
PROPERTIES OF ROBIN HOOD

- At least one record is in its home bucket ([illegible])

- If a fixed rule is used to break ties (eg: lexicographical order), then the final table is the same regardless of the order of insertion.

  [• We may search for the last element inserted (x_{n+1}) with $h(x_{n+1}) = 0$]

- The keys are stored in nondecreasing order by hash value, starting at some location k and wrapping around (From key 10)



In the first L buckets we have two kind of records:
- Records that overflow
- Records that hash to 0 and won ties

- We study just the overflow area (Linear Probing Sort) and then the search cost. (Gonnet-Munro 1984, Carlson-Poblete-Munro 1987, Viola-Poblete 1997)

POISSON TRANSFORM

- $\mathcal{S}_m[f_{m,n}\,;\, b\alpha] = e^{-mb\alpha} \sum_{n\geq 0} \frac{(bm\alpha)^n}{n!} f_{m,n} = \sum_{k\geq 0} a_{m,k} (b\alpha)^k$
- $f_{m,n} = \sum_{k\geq 0} a_{m,k} \frac{n^{\underline{k}}}{m^k}$

DISTRIBUTION OF BUCKET OCCUPANCY

- $Q_{m,n,d} \rightarrow$ # of ways to insert $\underline{n}$ records in a table with $\underline{m}$ buckets of size $\underline{b}$ so that a given bucket has more than $\underline{d}$ empty slots.

$$\frac{Q_{m,n,b-d+1}}{m^n} - \frac{Q_{m,n,b-d}}{m^n} = \Pr\{\text{given bucket has exactly } \underline{d} \text{ elements}\}$$

- $Q_{0,n,d} = [n=0]$
- $$Q_{m,n,d} = \begin{cases} \sum_{j=0}^{n} \binom{n}{j} Q_{m-1,j,d} & (0\leq n < bm-d) \\ 0 & (n \geq bm+d) \end{cases}$$

- If $[Q_{m,d}(z)]_n = \sum_{k\geq 0}^{n} Q_{m,k,d} \frac{z^k}{k!}$ we have

$$\begin{cases} Q_{0,d}(z) = 1 & \\ Q_{m,d}(z) = \underbrace{[e^z Q_{m-1,d}(z)]_{bm-d-1}} & (m\geq 1) \end{cases}$$

Key idea! ⟶ New sequence of numbers

- If we do **NOT** have truncation then

$$\left.\begin{cases} Q_{0,d}(z) = 1 \\ Q_{m,d}(z) = e^{z} Q_{m-1}(z) \quad (m > 0) \end{cases}\right\} \Rightarrow Q_{m,d}(z) = e^{mz}$$

Truncation complicates the solution and it does not seem to be a "nice" closed formula for $Q_{m,d}(z)$.

KEY IDEA! Find a sequence of numbers $T_{k,d,b}$ with $k \geq 0$

and $T_{d,b}(z) = \sum_{k \geq 0} T_{k,d,b} \frac{z^k}{k!}$, INDEPENDENT of m

such that $\boxed{Q_{m,d}(z) = \left[T_{d,b}(z) e^{mz}\right]_{bm-d-1}}$

Moreover satisfies $\boxed{\lim_{m \to \infty} \mathcal{P}_m \left[\frac{Q_{m,n,d}}{m^n}, b\alpha\right] = T_{d,b}(b\alpha)}$

The family $T_d(b\alpha)$ with $0 \leq d \leq b-1$ captures the distribution of bucket occupancy in the Poisson Model!!!

The family of sequences $T_{k,d,b}$

- $T_{k,0,1} \longrightarrow 1, -1, 0, 0, 0, 0, \cdots$

- $T_{k,0,2} \longrightarrow 1, 0, -1, 2, -8, 48, -378, 3672, -42368, \cdots$
- $T_{k,1,2} \longrightarrow 1, -1, 1, -2, 8, -48, 378, -3672, 42368, \cdots$

- $T_{k,0,3} \longrightarrow 1, 0, 0, -1, 3, -6, -12, 264, -2016, \cdots$
- $T_{k,1,3} \longrightarrow 1, 0, -1, 2, -3, -2, 60, -408, 1341, \cdots$
- $T_{k,2,3} \longrightarrow 1, -1, 1, -1, 0, 8, -48, 144, 675, \cdots$

- These sequences do not appear in Sloane's Encyclopedia.

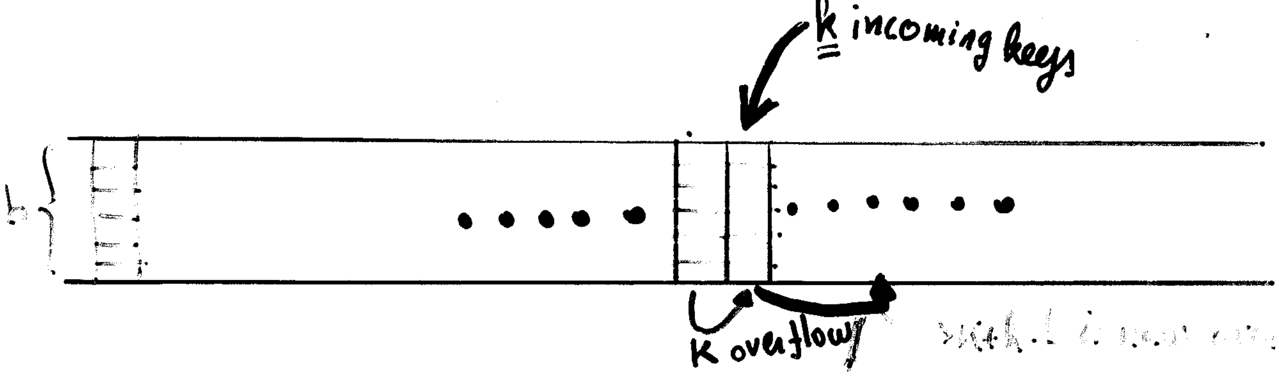
Some interesting properties

- $\displaystyle\sum_{j=0}^{k} \binom{k}{j} \left(\left\lfloor \frac{k+d}{b} \right\rfloor \right)^{k-j} T_{j,d,b} = [k=0]$ → implicit definition!
- $$\sum_{d=0}^{b-1} T_{k,d,b} = \begin{cases} b & (k=0) \\ -1 & (k=1) \\ 0 & (k>1) \end{cases} \Rightarrow \sum_{d=0}^{b-1} T_{d,b}(bx) = b(1-x)$$
- $T_{0,d,b} = 1 \qquad T_{k,d,b} = 0$ if $1 \le k \le b-d-1$
- Blake & Konheim (1977) present $T_{0,b}(x)$ and gives some properties.
- Much more work to be done with $T_{k,d,b}$!!!

[illegible]

Poisson Analysis of the overflow area

- We have $m, n \to \infty$ and $\frac{n}{bm} = \alpha \quad (0 \le \alpha < 1)$
- $\Pr[k \text{ keys hash to a given bucket}] = e^{-b\alpha} \frac{(b\alpha)^k}{k!}$



$\Omega(b\alpha, z) \to$ PGF for # elements that overflow from a bucket.

First approach: $\Omega(b\alpha, z) \sim \frac{\Omega(b\alpha, z)\, e^{b\alpha(z-1)}}{z^b}$ <u>NOT</u> valid if bucket receives less than <u>b elements</u>

Correction term: $\sum_{s=1}^{b} (1 - z^{-s})\, P_s(b\alpha)$ → Pr. [having ≤ s elements in bucket]

$P_s(b\alpha) = T_{s-1}(b\alpha) - T_s(b\alpha)$

After calculations: $\boxed{\Omega(b\alpha, z) = \frac{z-1}{z^b - e^{b\alpha(z-1)}} \sum_{s=0}^{b-1} T_s(b\alpha)\, z^{-s}}$

→ INDEPENDENT of the hash function used!!

When $b=1$ $\Omega(\alpha, z) = \frac{z-1}{z - e^{\alpha(z-1)}}(1-\alpha)$

$\frac{z-1}{z - e^{\alpha(z-1)}}$ → eulerian numbers!

("easy" combinatorial interpretation?)

P.G.F FOR THE SEARCH COST

- We search for an element pushed by
  - overflow from previous bucket
  - other elements hashed in same bucket

- Let $C(b\alpha, z) = \sum_{k\geq 0} c_k(b\alpha) z^k$ the P.G.F. for the total displacement without considering the fact that we should only count # buckets.

So, the P.G.F. we need is

$$\Upsilon(b\alpha, z) = z \sum_{k\geq 0} c_k(b\alpha) z^{\lfloor k/b \rfloor} = \sum_{k\geq 0} \left(\sum_{d=0}^{b-1} c_{bk+d}(b\alpha) \right) z^k$$

$$\boxed{\Upsilon(b\alpha, z) = \frac{z}{b} \sum_{j=0}^{b-1} C(b\alpha, e^{\frac{2\pi i}{b} j} z^{1/b}) \sum_{p=0}^{b-1} (e^{\frac{2\pi i}{b} j} z^{1/b})^{-p}}$$

- If $\underline{k}$ elements collide with the searched one, the total displacement is: (expected) originated by separately retrieving all these records is:

$$\frac{1}{k+1} \sum_{r=0}^{k} z^r = \frac{1}{k+1} \frac{1-z^{k+1}}{1-z}$$

with P.G.F

$$\frac{e^{-b\alpha}}{1-z} \sum_{k\geq 0} \frac{(b\alpha)^k}{(k+1)!} (1-z^{k+1}) = \frac{1-e^{b\alpha(z-1)}}{b\alpha(1-z)}$$

- Then, we have

$$1-e^{b\alpha(z-1)} \qquad e^{b\alpha(z-1)} - 1 \qquad \sum^{b-1} T_{b-1-s}(b\alpha) z^s$$

DISTRIBUTION OF SEARCH COST

Let $T_{b\alpha}$ the R.V. for the cost of searching a random element in a $b\alpha$-full table with buckets of size $\underline{b}$ using the Robin Hood linear probing algorithm.

$$\boxed{T_n(b\alpha) = \Pr\{T_{b\alpha} = n+1\} = \sum_{s=0}^{b-1} \frac{T_{b-1-s}(b\alpha)}{b\alpha} \sum_{k\geq 1} e^{-bk\alpha} \sum_{d=0}^{b-1} \left(\frac{(kb\alpha)^{b(n+k)+d-s}}{(b(n+k)+d-s)!} - \frac{(kb\alpha)^{b(n+k+1)+d-s}}{(b(n+k+1)+d-s)!} \right)}$$

- For fixed m, n the result follows from depoissonization. (EXACT!)

First moment

- $$E[T_{b,m,n+1}] = 1 + \sum_{k=1}^{\lfloor n/b \rfloor} \sum_{i=kb}^{n} (-1)^{i-kb} \binom{i-1}{kb-1} \frac{(kb)^i}{(i+1)!} \frac{n^{\underline{i}}}{(bm)^i}$$

- $$E[T_{b\alpha}] = 1 + \sum_{k\geq 1} e^{-bk\alpha} \sum_{n\geq 1} n \frac{(bk\alpha)^{n+bk-1}}{(n+bk)!}$$

- $$bE[T_{b,m,bm-1}] = \frac{\sqrt{2\pi bm}}{4} + \frac{1}{3} + \sum_{d=1}^{b-1} \frac{1}{1 - T\left(e^{\frac{2\pi i}{b}d-1}\right)} + \frac{1}{48}\sqrt{\frac{2\pi}{bm}} + O\left(\frac{1}{bm}\right)$$

HIGHER MOMENTS

$$E\left[\mathcal{T}_{b\alpha}^{\,r}\right] = r\sum_{s=0}^{b-1}\frac{T_{b-1-s}(b\alpha)}{b\alpha}\sum_{n\geq 1} n^{r-1}\sum_{k\geq 1} e^{-kb\alpha}\sum_{d=0}^{b-1}\frac{(bk\alpha)^{b(k+n)+d-s}}{(b(k+n)+d-s)!}$$

$$+\,1^{r}\sum_{s=0}^{b-1}\frac{T_{b-1-s}(b\alpha)}{b\alpha}\sum_{k\geq 1} e^{-bk\alpha}\sum_{d=0}^{b-1}\frac{(bk\alpha)^{bk+d-s}}{(bk+d-s)!}$$

- For exact m, n we do depoissonization.

Limit distributions

- When $\alpha\to 1$ main asymptotic contribution of $E[\mathcal{T}_{b\alpha}^{\,r}]$ comes from r^{th} derivatives of $C(b\alpha, z^{1/b})$ in $\mathcal{T}(b\alpha, z)$.
- As a consequence, proper generalizations of limit distributions for $b=1$ apply:

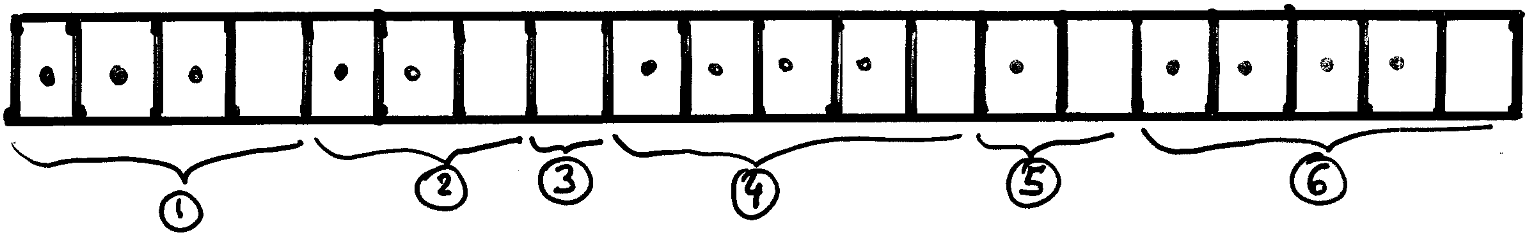
- When $\alpha\to 1$ $\quad b\Pr\{(1-\alpha)\mathcal{T}_{b\alpha}\leq x\}\to$ Exponential $(b/2)$
- When $m\to\infty$ $\quad b\Pr\left\{\frac{\mathcal{T}_{m,bm-1}}{\sqrt{bm}}\leq x\right\}\to$ Rayleigh $(1/2)$

TOTAL CONSTRUCTION COST

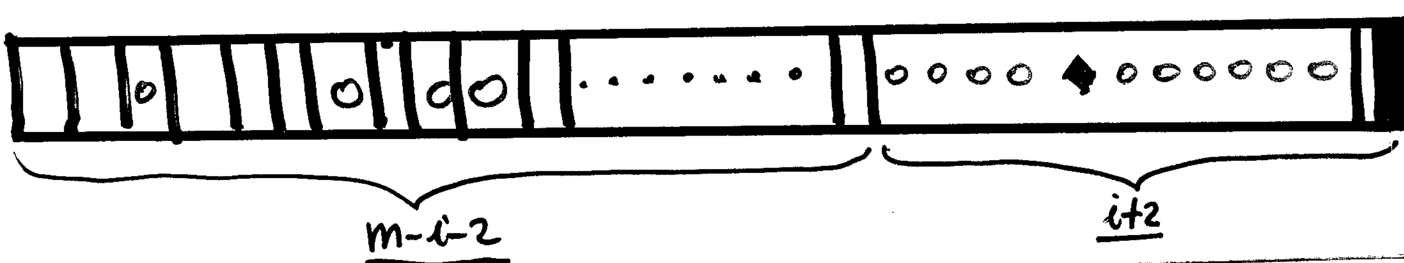
(Knuth, Flajolet-Poblete-Viola 1997) → $b=1$

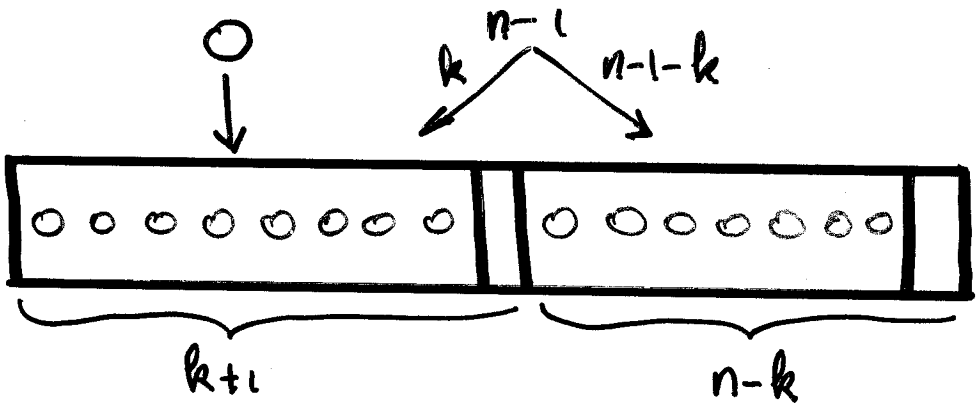
COMBINATORIAL INTERPRETATION

- A linear probing hash table is a sequence of "almost full tables" (clusters).



- SO, ENOUGH TO STUDY ALMOST FULL TABLES!
- (Monroe, Poblete, Viola 1997) → The Poisson transform applied to a R.V. in a full table, is the same as the Diagonal Poisson Transform applied to the same RV studied in an almost full table !!! → For general b?





- $F_{n,k} \to$ # ways of creating an AFT with n elements and total displacement k

- $$F(z,q) = \sum_{n,k\geq 0} F_{n,k}\, q^k \frac{z^n}{n!}$$

- $$F_n(q) = \sum_{k=0}^{n-1} \binom{n-1}{k} F_k(q)\,(1+q+\cdots+q^k)\, F_{n-1-k}(q)$$

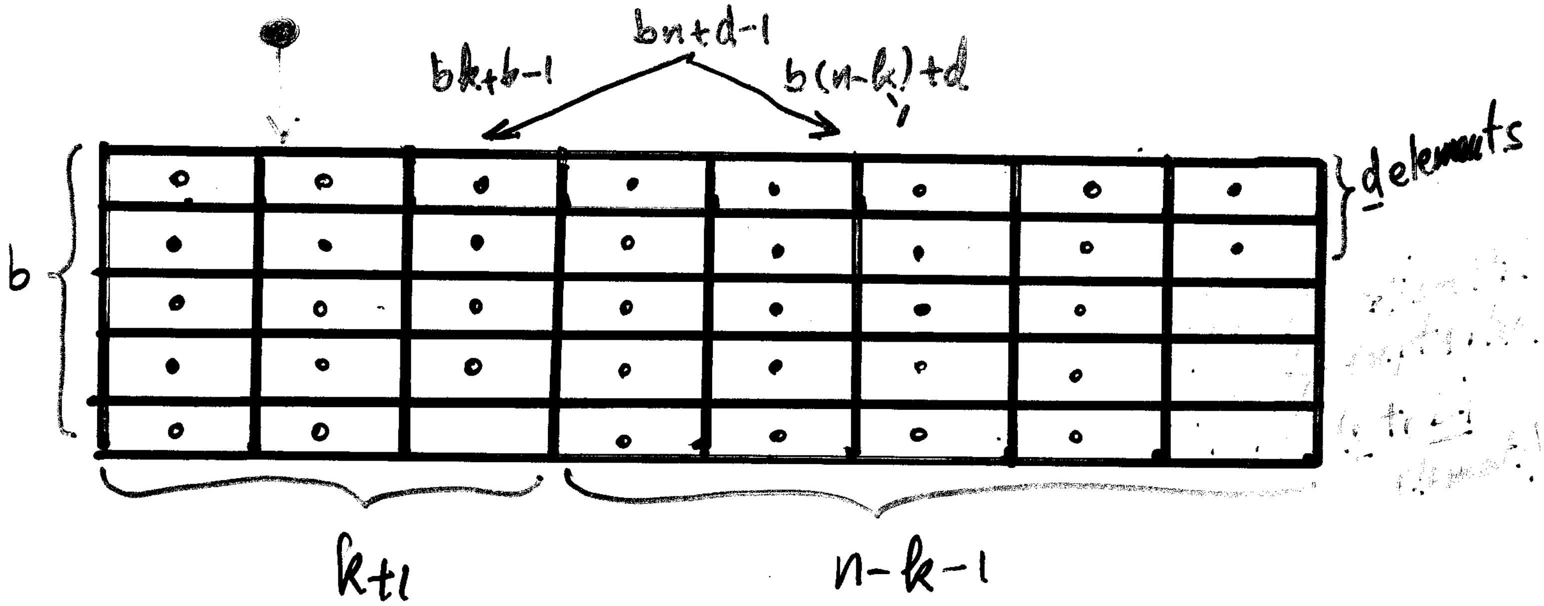
$$\boxed{\frac{\partial}{\partial z} F(z,q) = F(z,q)\, \frac{F(z,q) - q\,F(zq,q)}{1-q}}$$

$\to HF(z,q)$

- If $d_{n,n-1}$ is the R.V. for the total displacement in a AFT then

$$\Pr\left\{ \frac{d_{n,n-1}}{\left(\frac{n}{2}\right)^{3/2}} \leq x \right\} \xrightarrow[(n\to\infty)]{} \Pr\{X \leq x\}$$

$X \to$ Airy



- $F_d(q) = 1$
- $F_{bn+d}(q) = \sum_{k=0}^{n-1} \binom{bn+d-1}{bk+d-1} F_{bk+b-1}(q) \frac{1-q^{k+1}}{1-q} F_{b(n-k-1)+d}(q)$

$$+ [d>0] \frac{1-q^{n+1}}{1-q} F_{bn+d-1}(q)$$

- $F_d(z,q) = \sum_{n\geq 0} F_{bn+d}(q) \frac{z^{bn+d}}{(bn+d)!}$

- $HF_d(z,q) = \sum_{n\geq 0} \frac{1-q^{n+1}}{1-q} F_{bn+d}(q) \frac{z^{bn+d}}{(bn+d)!}$

$$\boxed{F_0(z,q) = e^{\int_0^z H F_{b-1}(t,q)\,dt}}$$

$$\boxed{F_d(z,q) = F_0(z,q) \int_0^z \frac{HF_{d-1}(t,q)}{F_0(t,q)}\,dt \quad (1 \leq d < b)}$$

[→ Solve it!!! → limit distributions? → not even moments!

INDIVIDUAL DISPLACEMENTS: FCFS ($b \geq 1$)

– ([illegible])

- Very interesting ideas and results
- Extremely difficult to read!!
- It does NOT use combinatorial interpretation of linear probing
  - ↳ NEW results at the light of interpreting combinatorially the results in the paper!!
- $$T_{b,d}(z) = \sum_{n \geq 0} F_{bn+d} \frac{z^n}{(bn+d)!} = \frac{1}{z^{d/b}} F_d(z^{1/b}, 1)$$
  ↳ it does NOT use q-analogs
- Given a sequence $\{\alpha_k^{(z)}, 1 \leq k \leq n\}$ the elementary symmetric functions $\{\sigma_k(z)\}$ of the sequence $\{\alpha_k(z)\}$ are defined as the coefficients of the polynomial $$\sum_{k=0}^{n} \sigma_k(z)\, x^{n-k} = \prod_{k=1}^{n} (x + \alpha_k(z))$$
- Given $\boxed{\{\alpha_d(z) = T(r^d z) : 0 \leq d \leq b-1\}}$ with $T \to$ tree function, $r = e^{\frac{2\pi i}{b}} \to b^{\text{th}}$ root of unity

  we have $(bz)^b\, T_{b,d}((bz)^b) = (-1)^{b-d-1}\, b^{b-d}\, \sigma_{b-d}(z)$

  $$\boxed{z^{b-d} F_d(bz, 1) = (-1)^{b-d-1} \sigma_{b-d}(z)}$$

  ↳ CAN it be generalized for q-analogs of F_d??

[illegible]

$$\cdot \Lambda(z,w) = \sum_{m\geq 0} \left(\sum_{n=0}^{bm-1} Q_{m,n,0} \frac{z^n}{n!} \right) w^{bm}$$

· After lengthy and complicated derivation

$$\Lambda(z,w) = \frac{N(z,w)}{1-N(z,w)} \quad \text{with } N(z,w) = \sum_{d=0}^{b-1} w^{b-d} F_d(wz,1)$$

$$= 1 - \prod_{d=0}^{b-1} \left(1 - \frac{b}{z} T\left(r^d \frac{wz}{b}\right)\right) \quad *$$

"TRIVIAL" proof if we use combinatorial interpretation!

· If $w=1$ * translates into $\boxed{\sum_{n\geq 0} F_n \frac{z^n}{n!} = 1 - \prod_{d=0}^{b-1} \left(1 - \frac{b}{z} T\left(r^d \frac{z}{b}\right)\right)}$

$\hookrightarrow$ q-analog?

• NEW!!!

by combinatorial interpretation!

$$\boxed{\Lambda_d(z,w) = \sum_{m\geq 0} \left(\sum_{n=0}^{bm-d-1} Q_{m,n,d} \frac{z^n}{n!} \right) w^{bm}}$$

$$\boxed{\Lambda_d(z,w) = \frac{N_d(z,w)}{1-N(z,w)}}$$

$$\boxed{N_d(z,w) = N(z,w) - \sum_{j=b-d}^{b-1} \frac{1}{z^{b-j}} [x^j] \left(x^b - bT\left(r^j \frac{zw}{b}\right)\right)}$$

Displacement of n^{th} ball

• $G_{m,n,j}$ = # ways of inserting the n^{th} ball with displacement j

$$G(z,w,q) = \sum_{m \geq 0} w^{bm} \sum_{n=1}^{bm} \frac{(mz)^{n-1}}{(n-1)!} \sum_{j=0}^{n} G_{m,n,j}\, q^{j}$$

after lengthy derivation!

$$= \Lambda(z,w) \underbrace{\sum_{d=0}^{b-1} w^{b-d}\, HF_d(wz,1)}_{\text{insert in last bucket!}}$$

Have any hash table with last bucket not full

"TRIVIAL" with combinatorial interpretation!

Generating Function of $T_0(b\alpha)$ ← !!!

$$\lim_{\substack{m,n \to \infty \\ \frac{n}{m} = b\alpha}} \frac{Q_{m,n,0}}{m^n} = \boxed{T_0(b\alpha) = \frac{b(1-\alpha)}{\prod_{d=1}^{b-1}\left(1 - \frac{T(r^d \alpha e^{-\alpha})}{\alpha}\right)}}$$

they did not realize that!

→ for general $T_d(b\alpha)$? → can use implicit definition of $T_{k,d}$?

Expected cost of a random element

• $E[C(b\alpha)] \sim \frac{1}{2b(1-\alpha)}$ → give asymptotic but not exact value.

CONCLUSIONS AND FUTURE WORK

- We present the first distributional analysis of a linear probing hashing algorithm with buckets

- Key component of the analysis is the introduction of a new family of sequences $T_{k,d,b}$

- Better understanding of the properties of $T_{k,d,b}$, like combinatorial identities, asymptotic expansions. $\Rightarrow$ complete understanding of occupancy distributions in buckets

- Combinatorial relation with other problems like Random walks?

- Find other problems where $T_{k,d,b}$ may appear.