RECORDINALITY
les grands flux de données vus comme des permutations aléatoires

— 5 mars 2012, Journées Aléa, Luminy —

Jérémie Lumbroso
LIP6 / INRIA Rocquencourt

travail en cours réalisé avec

Ahmed Helmi
Univ. Politècnica de Catalunya

Conrado Martínez
Univ. Politècnica de Catalunya

Alfredo Viola
Instituto de Computación

1. SURVOL des estim. de cardinalité

- **échantillonnage** : intrinsèquement limité pour raisons statistiques, mais + de 100 réf., notamment en *biologie* [Bunge & Fitzpatrick 1993]
- **collectionneur de coupons** [Whang *et al.* 1990, Estan *et al.* 2003...]
- **statistiques d'ordre**
  - Probabilistic Counting [FlMa85]: $\min(\mathbb{N} \setminus \{G_1, \ldots, G_n\})$
  - variables uniformes [Bar-Yossef *et al.* 2002, Giroire 2005...]: $X_{(k)}, k > 2$
  - LogLog, etc. [FlDu03, FlFuGaMe07]: $\max\{G_1, \ldots, G_n\}$
  - estimation par le minimum [L. 2010]: $\min\{X_1, \ldots, X_2\}$

où

$$G_i \in \mathrm{Geo}(1/2) \qquad X_i \in \mathcal{U}(0,1) \qquad X_{(k)} \in \mathrm{Beta}(k, n-k+1)$$

Démarche générique (dans l'ex. $N = 8, n = 6$)

- un flux — Allons mon pauvre coeur allons mon vieux complice
- hachage (v.a. en Post-it®) — $\underset{0.26}{\text{Allons}}\ \underset{0.58}{\text{mon}}\ \underset{0.22}{\text{pauvre}}\ \underset{0.68}{\text{coeur}}\ \underset{0.26}{\text{allons}}\ \underset{0.58}{\text{mon}}\ \underset{0.52}{\text{vieux}}\ \underset{0.14}{\text{complice}}$
- fonction insensible aux répétions — $\min = \mathbf{0.26}, 0.26, \mathbf{0.22}, 0.22, 0.22, 0.22, 0.22, \mathbf{0.24}$
- ... dont l'espérance dépend de n

$(\mathbb{E}_n[F] = f(n) \quad \Rightarrow \quad \mathbb{E}_n\left[f^{-1}(F)\right] \approx n)$ $\quad \mathbb{E}_n[\min] = 1/(n+1), \qquad 1/0.14 - 1 \approx 6$

---

ces techniques utilisent **l'ultime valeur calculée =**
dépend uniquement de la composition (+ fonction hachage)

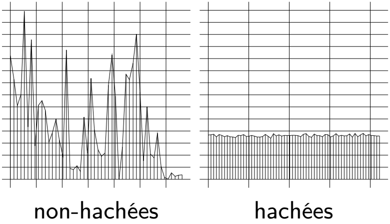
Ex. : statistiques d'ordre k d'uniformes [Bar-Yossef *et al.* 2002]

La théorie probabiliste (élémentaire) :

- tirer n variables uniformes, $X_i \in \mathcal{U}(0,1)$
- $X_{(k)}$ est la k-ième plus petite valeur ($k = 1$, min. ; $k = n$, max.), i.e.

$$\left|\left\{i, X_i \leqslant X_{(k)}\right\}\right| = k$$

- **propriétés connues :**
  - $X_{(k)} \in \mathrm{Beta}(k, n-k+1)$
  - $\mathbb{E}\left[X_{(k)}\right] = k/(n+1)$
  - pour $k > 2$, $\mathbb{E}\left[k/X_{(k)}\right] \approx n$



L'algorithmique (élégante) :

- lire chaque élément $x \in \mathcal{A}^*$ du flux
- hacher x pour obtenir $h(x) \in \mathcal{U}(0,1)$
- tenir tableau T contenant k plus petites valeurs hachées (triées)
- à tout moment $k/T[k]$ estime le nombre d'éléments distincts vus

des flux aux permutations...

- avec hachage, qui associe à tout élément x distinct, $h(x) \in \mathcal{U}(0,1)$:
  - chaque élément **distinct** est associé à une variable aléatoire uniforme
  - les variables aléatoires uniformes forment une **permutation aléatoire** (d'ailleurs, méthode standard pour générer celles-ci [Flaj. *et al.* 11...])
- **mais** les estimateurs existants ignorent cette permutation aléatoire

or il y a des informations contenues dans la permutation, qui permettent de faire des estimateurs probabilistes

Suite de l'exposé : trouver des statistiques sur les permutations

- dont l'espérance est fonction de la taille de la permutation (par ex.)
- qui sont faciles et peu couteuses à calculer

et en faire des algorithmes de comptage probabiliste.

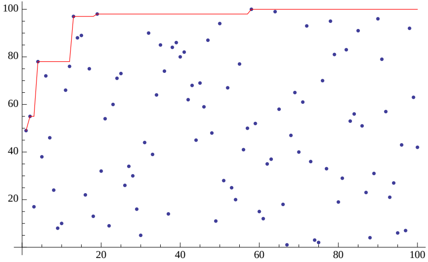
2. l'algo RECORDINALITY

Définition : soit σ une permutation, σ_i est dit *record* (ou *maximum partiel*) si

$$\forall j < i, \quad \sigma_j < \sigma_i.$$

Exemple : $\{\mathbf{1}, \mathbf{7}, \quad 3, \quad \mathbf{9}, 8, 4, 2, 5, \mathbf{10}, 6\}$, cette permutation a 4 records.

Résultat bien connu : si r est le nombre de records d'une permutation de taille n, alors $\mathbb{E}_n[r] \sim \log n$.



Algorithmiquement, si on lit une permutation de gauche à droite, tout en maintenant un registre contenant le maximum des éléments parcourus : nombre de records = nombre de modifications du registre.

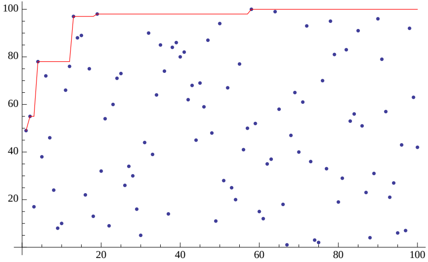
2. l'algo RECORDINALITY

Définition : soit σ une permutation, σ_i est dit *record* (ou *maximum partiel*) si

$$\forall j < i, \quad \sigma_j < \sigma_i.$$

Exemple : $\{\mathbf{1}, \mathbf{7}, 7, 3, 3, \mathbf{9}, 8, 4, 2, 5, \mathbf{10}, 6\}$, cette permutation a 4 records.

Résultat bien connu : si r est le nombre de records d'une permutation de taille n, alors $\mathbb{E}_n[r] \sim \log n$.



Algorithmiquement, si on lit une permutation de gauche à droite, tout en maintenant un registre contenant le maximum des éléments parcourus : nombre de records = nombre de modifications du registre.

extension standard aux k-records

Définition : soit σ une permutation, σ_i est dit *k-record* si

$$\#\{\sigma_j : j < i \text{ et } \sigma_j > \sigma_i\} < k.$$

Algorithmiquement, on lit la permutation séquentiellement de gauche à droite, en maintenant un tableau des k-plus grands éléments : nombre de k-records = nombre de modifications du tableau.

Proposition [ArMa09, HeMaPa12] : si r_k est le nombre de k-records alors

$$\mathbb{E}_n[r_k] = k(H_n - H_k + 1), \tag{1}$$

où H_k est la série harmonique ; de plus, on a la distribution limite

$$\frac{r_k - k(\log n - \log k + 1)}{\sqrt{k(\log n - \log k - 1 + k/n)}} \to \mathcal{N}(0,1). \tag{2}$$

Comme $\mathbb{E}_n[r_k] \sim k \log n + O(1)$, on imagine que

$$\boldsymbol{\mathcal{R}_k := \phi_k \exp(\alpha_k r_k)}$$

(avec α_k et ϕ_k des facteurs de correction) est un bon estimateur de n, le

Record + Cardinality = Recordinality

Théorème 1 [Helmi, L., Martínez, Viola 2012] : soit $r_k \in \mathbb{N}$ le nombre de k-records observés, alors l'estimateur $\mathcal{R}_k$ défini par

$$\mathcal{R}_k := \phi_k \exp(\alpha_k r_k)$$

avec les facteurs

$$\alpha_k := \sqrt{1 + 2/k} - 1 \qquad \text{and} \qquad \phi_k := k\mathrm{e}^{\alpha_k{}^2 k/2 - k\alpha_k}$$

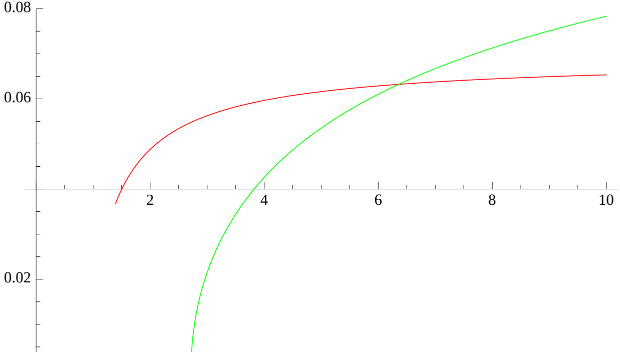
est un estimateur asymptotiquement non-biaisé du nombre n d'éléments distincts.

Théorème 2 [Helmi, L., Martínez, Viola 2012] : La précision de l'estimateur $\mathcal{R}_k$, donnée par l'erreur type, est :

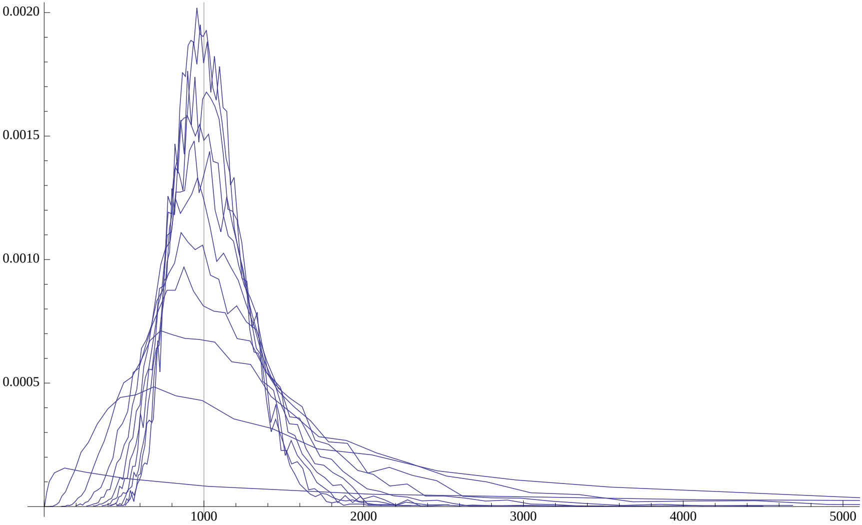
$$\mathbb{SE}_n[\mathcal{R}_k] = \frac{\sigma_n[\mathcal{R}_k]}{n} = \sqrt{\left(\frac{n}{k\mathrm{e}}\right)^{1/k} - 1}.$$

Particularité :

- précision se **dégrade** plus n est grand
- excellente précision quand n est moyen



Théorème 3: l'estimateur suit une loi log-normale.



distribution des estimations d'un flux de taille $n = 1000$, pour $k = 1, 5, 10, \ldots, 50$

A. $\boxed{\text{estimateurs HYBRIDES}}$

$$\mathcal{Z}_{\lambda,k} := \lambda \mathcal{O}_k + (1-\lambda)\mathcal{R}_k \qquad \lambda \in]0,1[$$

- **indépendance :** la valeur ultime prise par les k plus grands éléments ne dépend pas de leur ordonnancement dans le flux (et donc du nombre de k-records)
- **gain en précision :** $\sqrt{\mathbb{V}_n[\mathcal{Z}_{\lambda,k}]} = \sqrt{\lambda^2 \mathbb{V}_n[\mathcal{O}_k] + (1-\lambda)^2 \mathbb{V}_n[\mathcal{R}_k]}$, en pratique entre 2% et 15% supplémentaire par rapport à $\mathcal{O}_k$ seul
- **même complexité algorithmique :** $\mathcal{R}_k$ maintient même tableaux des k plus grands éléments que $\mathcal{O}_k$ + compteur $O(\log k + \log\log n)$
- **universalité :** peut être combiné avec n'importe quel estimateur !

B. échantillonnage distinct

Soit un flux de taille N (avec n éléments distincts)

$$\mathcal{S} = x_1\ x_2\ x_3\ \cdots\ x_N$$

- échantillonnage standard (chaque x_i pris avec prob. $1/N$), les éléments à faible fréquence sont noyés dans la masse : c'est le **problème** de l'aiguille dans une botte de foin
- **solution :** échantillonnage distinct [Flaj. 90, L. 2012 ; Gibbons 2001...], prendre chaque élément distinct avec prob. $1/n$

Exemple : échantillon de 1 élém. du flux $(1, 1, 1, 1, 2, 1, 1, \ldots, 1)$, $N = 1000$; avec échantillonnage standard, prob. $999/1000$ de prendre 1 et $1/1000$ de prendre 2 ; avec échantillonnage distinct, prob. $1/2$ de prendre 1 et prob. $1/2$ de prendre 2.

Proposition : lorsque l'algorithme RECORDINALITY a terminé, les k éléments (dont les posts-it/valeurs hachées sont) maximaux stockés sont **un échantillon uniforme de l'ensemble sous-jacent au flux.**

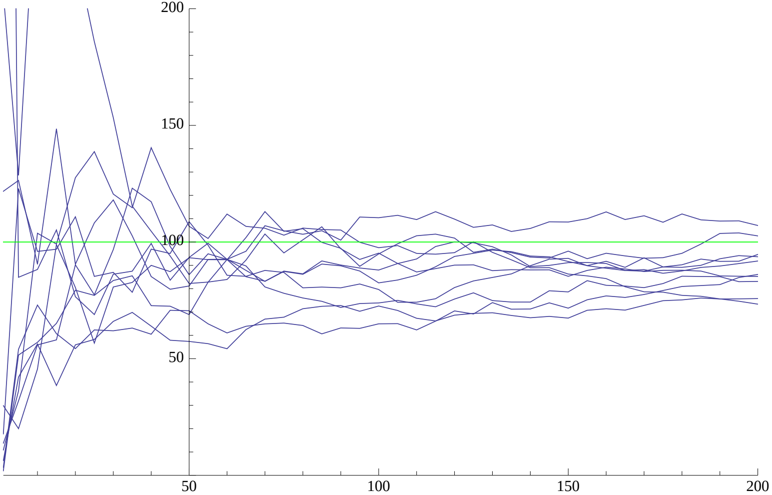
C. $\boxed{\text{a-t'on vraiment besoin de hachage ?!}}$

Précédents estimateurs (Prob. Counting, LogLog, etc.) utilisent les **valeurs de la fonction de hachage**, ils ont besoin de valeurs $x \in [0, 1]$.

RECORDINALITY ne se sert que du rang, de **l'ordre relatif des éléments** ; la fonction de hachage permute aléatoirement les données...

... mais on peut supposer que le flux est une permutation aléatoire (ou **suffisamment aléatoire**) et comparer directement les éléments :

- c'est le modèle théorique "*Random Stream*" [McGregor 2005...]
- appliquable en pratique ?
- comparaisons sur chaînes moins couteuses que faire hachage (qui représente 90% du temps de calcul dans algos existants) ?



précision des estimations en fonction du nombre k d'éléments retenus ;
(chaque courbe = une tragédie de Shakespeare ; ligne vert à 100% = repère)

3. $\boxed{\text{CONCLUSIONS}}$

- de nouveaux estimateurs basés sur les permutations aléatoires
- permettent d'obtenir des résultats complets à faible complexité en adaptant la littérature foisonnante existante
- semble exploiter davantage/différemment les données

perspectives :

- plein d'autres statistiques à étudier
- s'affranchir des fonctions de hachage ?
- algorithmes d'échantillonnage dont la taille de l'échantillon est fonction de n (au lieu d'être fixe)

Slides : `http://tinyurl.com/recordinality-alea2012-v1`