q-gram analysis and urn models

Pierre Nicodème

École polytechnique, Palaiseau
Laboratoire d'Informatique LIX

November 2003

Approximate pattern matching and the Jokinen-Ukkonen lemma

Def: q-gram any word of fixed size q

Edit operations over strings

- substitution $(l_1 \to l_2)$ $\quad aabdd \to aadcc$
- insertion $(| \to l)$ $\quad aa|dd \to aaecc$
- suppression $(l \to |)$ $\quad aaedd \to aa|cc$

Edit distance $\delta(S_1, S_2)$ between two strings S_1 and S_2

- minimum number of edit operations transforming S_1 into S_2

Jokinen-Ukkonen 1991 (loose version)

if $|S_1| = m$ and $\delta(S_1, S_2) \leq k$, then at least $m + 1 - (k + 1)q$ of the $m - q + 1$ q-grams of S_1 occur in S_2

Example

$$S_1 = aaabaaab$$

$$S_2 = aaacaaaa$$

$m = 8, \quad \delta(S_1, S_2) = 2 \rightarrow k = 2$

$$2 - \text{grams}(S_1) = \{\{aa, aa, ab, ba, aa, aa, ab\}\}$$

$$Q_{S_1,S_2} = 2 - \text{grams}(S_1) \text{ present in } S_2 = \{\{aa, aa, aa, aa\}\}$$

Jokinen-Ukkonen

$$|Q_{S_1,S_2}| \geq m + 1 - (k+1)q$$

$$4 \geq 8 + 1 - (2+1)2 = 3$$

Beware of the asymmetry: $|Q_{S_2,S_1}| = 5$

Application

When searching a pattern with errors in a text, slide over the text a window of same size as the pattern and discard windows which do not contain enough q-grams of the pattern

Aim of this work

Study of two statistics of q-grams in random sequences:

– number of " repeated" q-grams (number of q-grams occurring at least twice, without counting multiplicities

$$S = aaaabaaaabbb, \quad q = 2$$

$$Q_{\text{repeated}} = \{aa, ab, bb\} \quad |Q| = 3$$

– number of common q-grams to two sequences, without counting multiplicities

$$S_1 = aaaabaaaabbb$$

$$S_2 = aaaacaaaacbb$$

$$q = 2 \qquad Q_{\text{common}} = \{aa, bb\} \quad |Q| = 2$$

(Remark: symmetrical counting)

– Jokinen-Ukkonen statistics

Bernoulli non-uniform model for the sequences

A heuristic approach

Dependent model

FGSEWWTYURR . . . OOUYJREFDKB

FGSEWWTYU . . .
GSEWWTYU . . .
SEWWTYU . . .
EWWWTYU . . .
. . .

Independent model

TTG
GSE
UHI
ROY
. . .

sequence length $l = n + q - 1 \Rightarrow n$ q-grams

1. analyse the independent model
2. perform simulations for the dependent model and compare with the independent model

Repeated q-grams

Equivalent problems

Input: an alphabet Σ ($|\Sigma| = s$), an integer q, a random sequence S of size $n + q - 1$

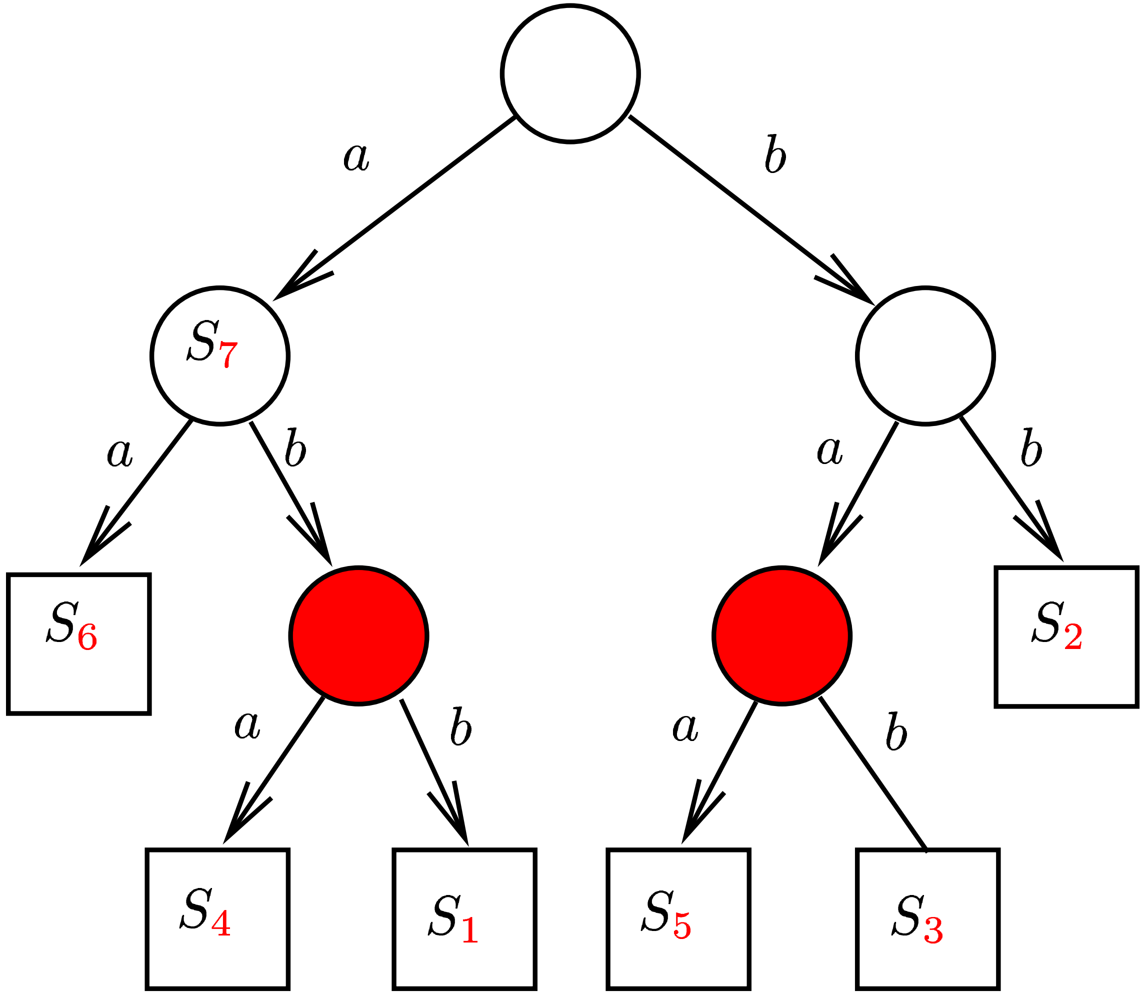
Dependent model

1. number of repeated q-grams
2. number of internal nodes at depth q of the suffix-tree build on S
3. number of self-intersections of a random walk of length n over the de Bruijn graph $B(s, q)$

Independent model

1. number of repeated q-grams
2. number of internal nodes at depth q of a trie build with n random keys over Σ
3. number of self-intersections of a random walk of length n over a complete graph $K(s^q)$
4. number of urns containing more than one ball in a system of s^q urns in which n balls are thrown

Suffix-trees



$S = abbabaa$ $\quad$ $q = 2$
1234567

$Q_{\text{repeated}} = \{ab, ba\}$

$|Q| = 2 =$ number of internal nodes at depth q

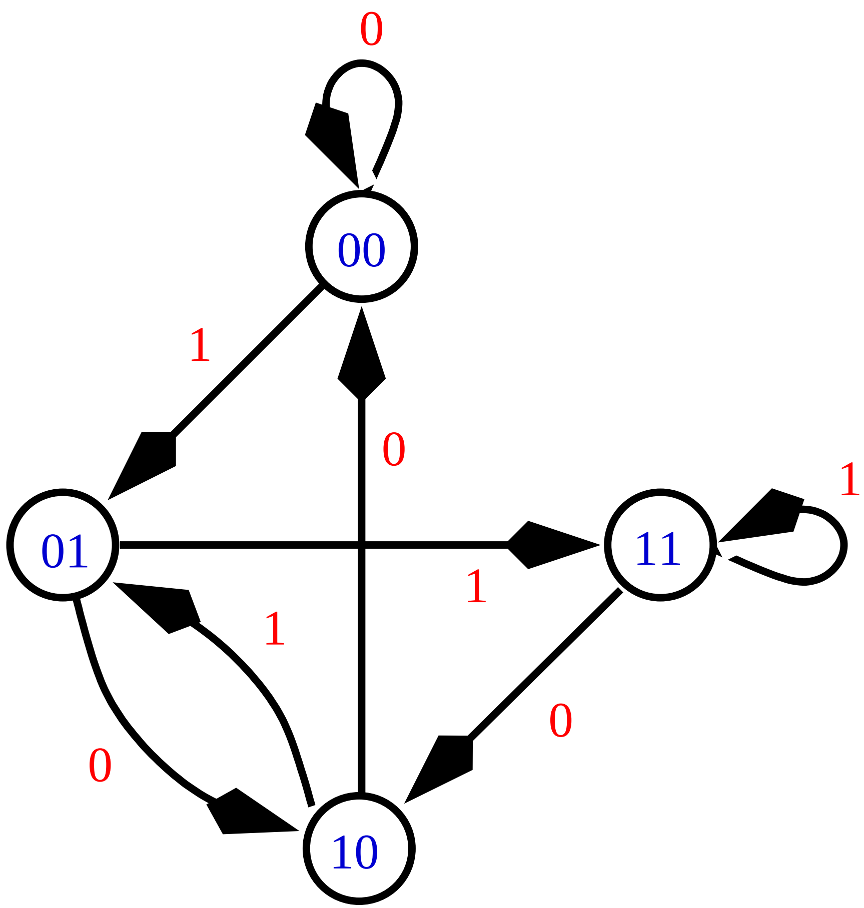
DE BRUIJN graphs

DE BRUIJN graph $B(s,q)$

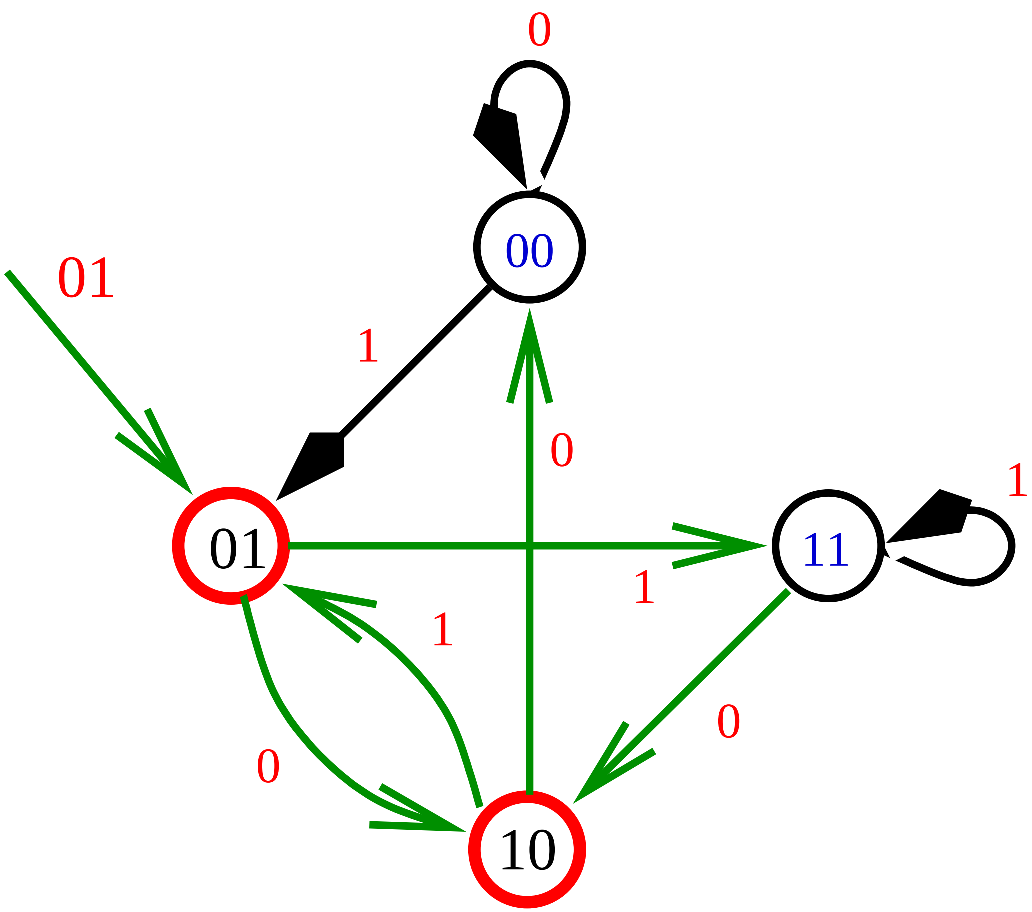
Vertices: $V = \{0, 1, 2, \ldots, s^q - 1\}$

Edges: $E = \{(v_i, v_j)\}$ with

$v_j = s \times v_i (\text{mod } s^q) + x, \quad x = 0|1|2|\ldots|(s-1)$



Random walks over de Bruijn graphs



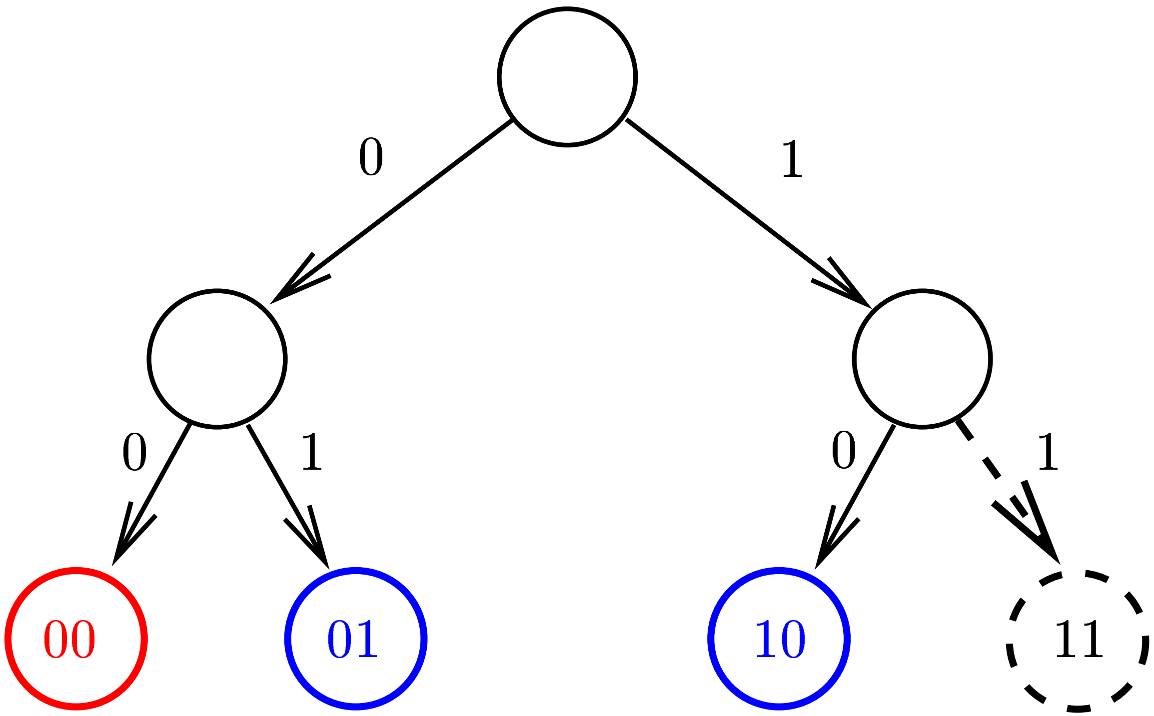
$S = abbabaa \qquad q = 2$

$S = 0110100$

$Q_{\text{repeated}} = \{01, 10\}$

$|Q| = 2 =$ number of vertices accessed more than once

Trie and urns (indep. model)



keys $= [00, 01, 00, 00, 10, 00]$ $Q_{\text{repeated}} = \{00\}$

$|Q| = 1 =$ number of nodes at depth q containing more than one key

equivalent to a system of 4 urns

key $\leftrightarrow$ number of urn

key $= w_{q-1}w_{q-2}\ldots w_0$

$\Pr(\text{urn}_i) = \Pr(\text{key}_i) = \prod_{0\le i\le q-1}\Pr(w_i)$

Previous results

– Guibas and Odlyzko - 1981, Rahman and Rivals - 2000, 2003
  enumeration of autocorrelations, missing words

– Szpankowski and Jacquet - 1994

  asymptotically, the distributions of path lengths of suffix-trees and of tries of same size are equal

  J. Fayolle - 2002, same result, but for the expectation

– Szpankowski and Sutinen - 1999

  phase transition in q-gram filtration

– urn models: numerous results

  Johnson and Kotz - 1977, Kolchin *et al.* - 1978, Drmota *et al.* - 2001, Flajolet *et al.* - 2003

Analysis of the urn model

X_n random variable counting the number of urns without collisions when n balls are thrown in the system of $m = s^q$ urns

$Y_n = m - X_n$ counts urns with collisions

G.F.

$$F(z, u) = \sum \Pr(X_n = k) u^k \frac{z^n}{n!}$$

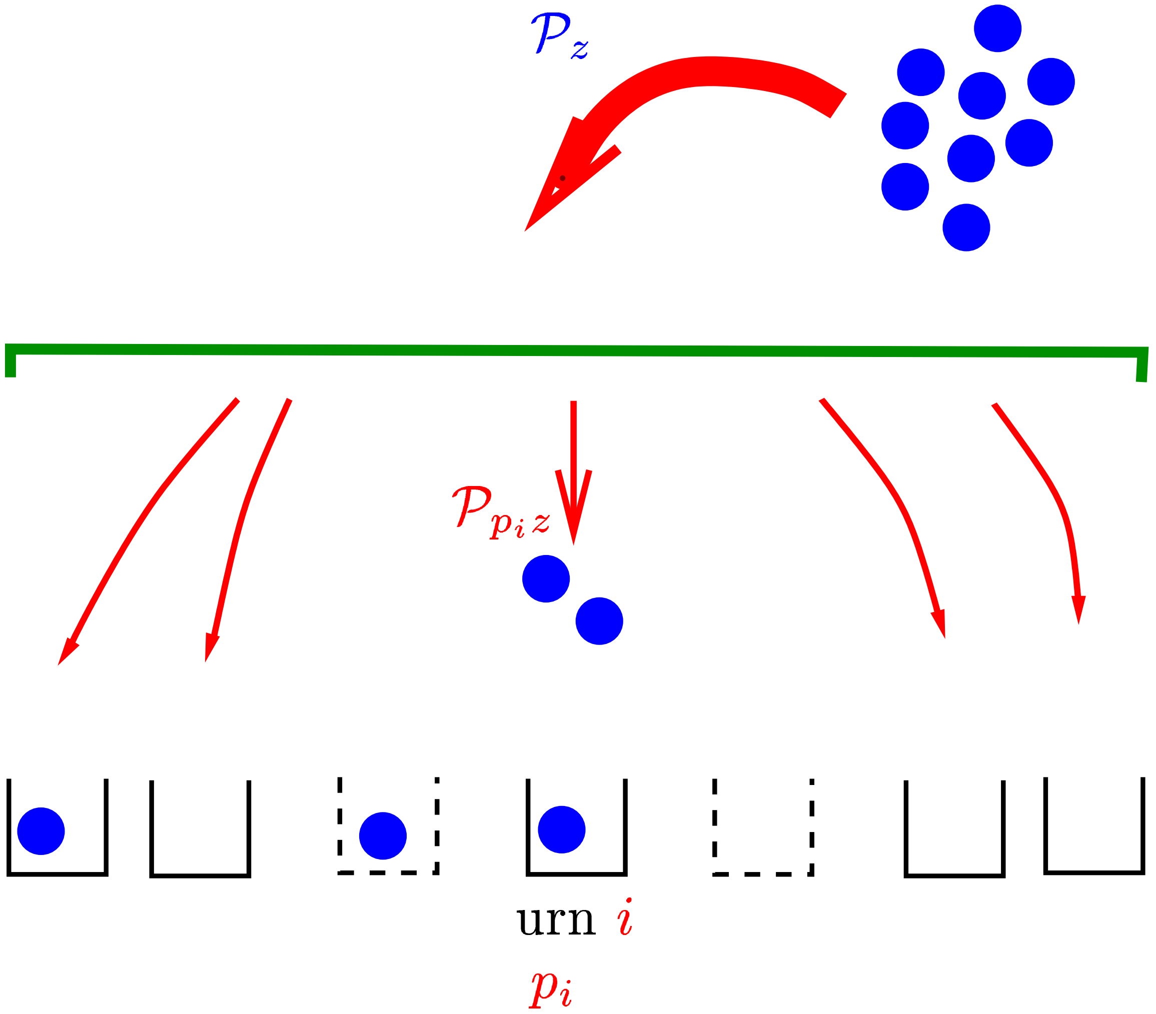
differentiations with respect to u

$\rightarrow$ gen. functions of moments of X_n

$\rightarrow$ extraction of nth Taylor coefficient and asymptotic evaluation

Poissonization

do not throw exactly n balls in the urns, but throw a random number of balls following a Poisson distribution.



The urns behave independently of each other

Poissonization - Depoissonization

$\mathcal{P}_{p_i z}$ balls in urn i.

$\Pr(\text{no collision}) = e^{-p_i z}(1 + p_i z)$.

u counts the urns without collisions

b.g.f. for urn i under the Poisson model

$\phi_i(z, u) = e^{-p_i z}\big((1 + p_i z)u + e^{p_i z} - 1 - p_i z\big)$

for the system of urns (Poisson again) $\Phi = \prod \phi_i$

$$\Phi(z, u) = e^{-z} \prod_{0 \le i \le m-1} (e^{p_i z} + (u - 1)(1 + p_i z))$$

$$\text{“exact” g.f.} \qquad F(z, u) = \sum f_n(u) \frac{z^n}{n!}$$

$$\Phi(z, u) = \sum_{n \ge 0} f_n(u) \frac{z^n}{n!} e^{-z} \Leftrightarrow f_n(u) = [z^n] n! e^z \Phi(z, u)$$

$$\Rightarrow F(z, u) = \prod_{0 \le i \le m-1} (e^{p_i z} + (u - 1)(1 + p_i z))$$

Expectation and standard dev.

$\mu_n = \mathbf{E}(X_n) \qquad m_n^{(2)} = \mathbf{E}(X_n^2)$

$$m(z) = \sum \mu_n z^n = \left.\frac{\partial F(z,u)}{\partial u}\right|_{u=1}$$

$$m^{(2)}(z) = \sum m_n^{(2)} z^n = \left.\frac{\partial}{\partial u} u \frac{\partial F(z,u)}{\partial u}\right|_{u=1}$$

extract $[z^n]m(z)$ and $[z^n]m^{(2)}(z)$ + asymptotics

when $n \times p_i \to \theta_i$

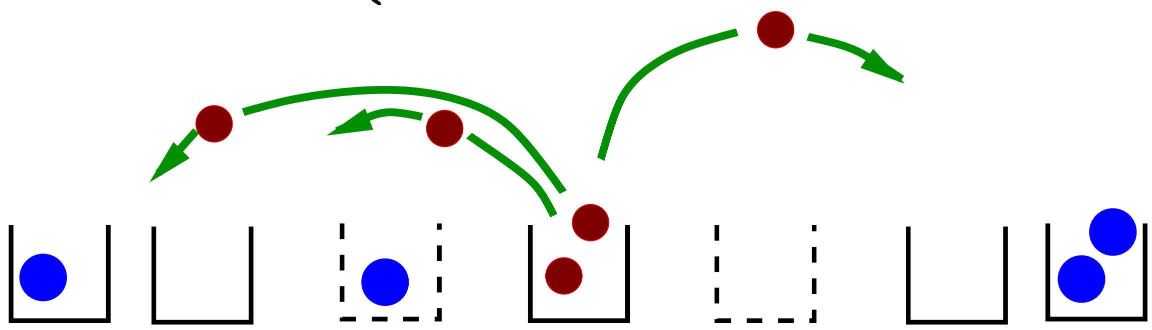
$$\mu_n = \sum_i \left(e^{-\theta_i}(1+\theta_i) + \frac{1}{2n} e^{-\theta_i}\theta_i^2(1-\theta_i) + O\left(\frac{1}{n^2}\right)\right) \quad \text{and} \quad \gamma_n = m - \mu_n$$

$$\sigma_n^2 = m_n^{(2)} - \mu_n^2 \approx \sum_i e^{-\theta_i}(1+\theta_i)\left(1 - e^{-\theta_i}(1+\theta_i)\right) \quad -\frac{1}{n}\left(\sum_i \theta_i^2 e^{-\theta_i}\right)^2$$

Poisson convergence (Chen-Stein)

number of empty urns: Barbour - Holst 1989

$$I_k = \begin{cases} 1 \text{ if urn } k \text{ empty} \\ 0 \text{ elsewhere} \end{cases} \qquad W = \sum_k I_k \quad \mu = \mathbf{E}(W)$$



(1) empty urn k by throwing the balls into the other urns

(2) coupling: after this operation

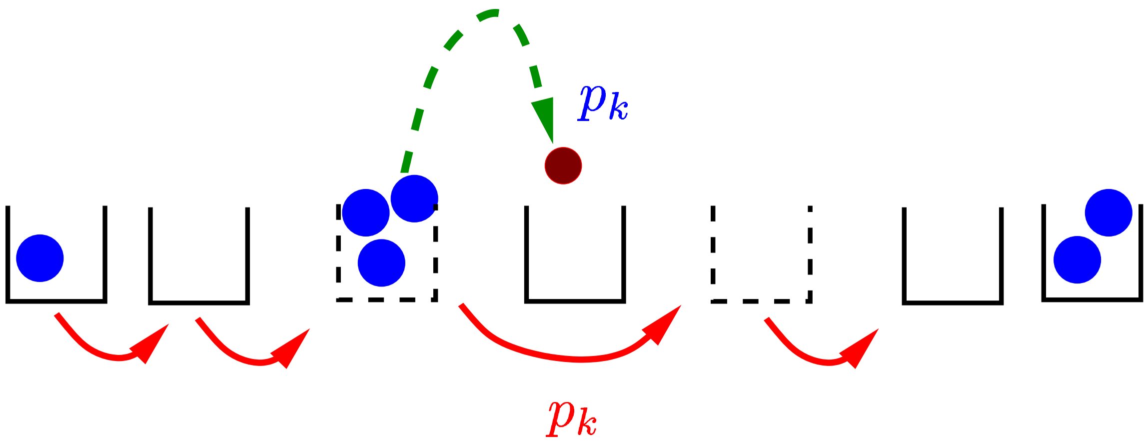
$$\begin{cases} J_{ik} = \begin{cases} 1 \text{ if urn } i \text{ empty} \\ 0 \text{ elsewhere} \end{cases} \\ I_{ik} = I_i \ \forall k \end{cases} \Bigg| \Rightarrow J_{ik} \leq I_{ik} \ (i \neq k)$$

$$\mathcal{L}(J_{1k}, \ldots, J_{mk}) = \mathcal{L}(I_{1k}, \ldots, I_{mk} \,|\, I_k = 1)$$

$$\Rightarrow d(W, \mathcal{P}_\mu) \leq \min(1, \mu) \left(1 - \frac{\mathbf{Var} W}{\mu}\right)$$

Poisson convergence (r-collisions)

$$I_k = \begin{cases} 1 \text{ if } \geq r \text{ balls in urn } k \\ 0 \text{ elsewhere} \end{cases} \quad W = \sum_k I_k \quad \mu = \mathbf{E}(W)$$



if less than r balls in urn k

repeat until there are $\geq r$ balls in urn k
  for all urns $i \neq k$
    for each ball in urn i
      throw it into urn k with proba. p_k

number of iterations finite with proba. 1

coupling + same proof as Barbour and Holst

$$\Rightarrow d(W, \mathcal{P}_\mu) \leq \min(1, \mu)\left(1 - \frac{\mathbf{Var}W}{\mu}\right)$$

Dependent model

Th: the language of words containing e repeated q-grams is rational, for all e

1. consider the DE BRUIJN directed graph $B(s,q)$ as an automaton $(\Sigma, Q, 0, \delta, F = Q)$ where the states of Q (vertices) are naturally numbered from 0 to $s^q - 1$ and all states are terminal

2. Consider 3^{s^q} copies of $B(s,q)$ corresponding of all combinations of labelling with $\lambda = 0|1|2$ of the vertices of $B(s,q)$

3. Number the copies along the numbering of the states and the labels:

   $B_N(s,q) \Leftrightarrow$ label of vertex n is the nth digit of N in base 3.

4. build a (huge) automaton $(\Sigma, \mathcal{Q}, 0_0, \Delta, \mathcal{Q})$

   where $\mathcal{Q} = \{0, 1, \ldots, s^q - 1\} \times \{0, 1, \ldots, 3^{s^q} - 1\}$ (notation $[n, N]$)

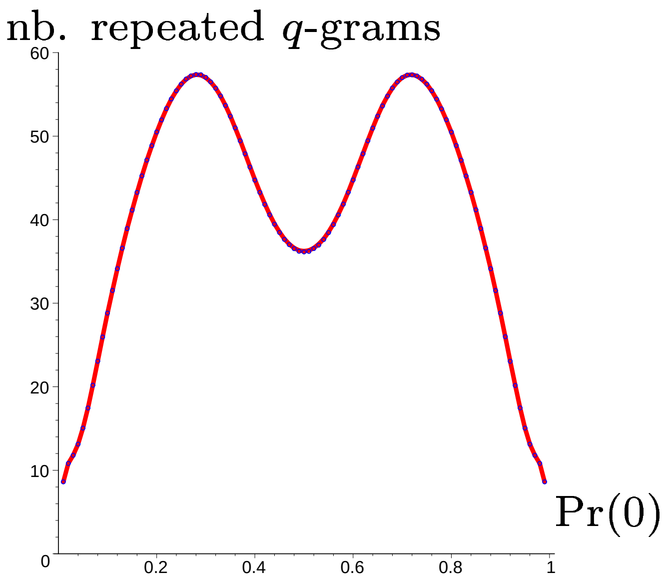
   by connecting the copies

$$\begin{cases} \lambda = 0,1 : \Delta([n, N_1], l) = [\delta(n, l), N_2] \ (N_2 = N_1 + 3^n) \\ \lambda = \quad 2 : \Delta([n, N_1], l) = [\delta(n, l), N_1] \end{cases}$$

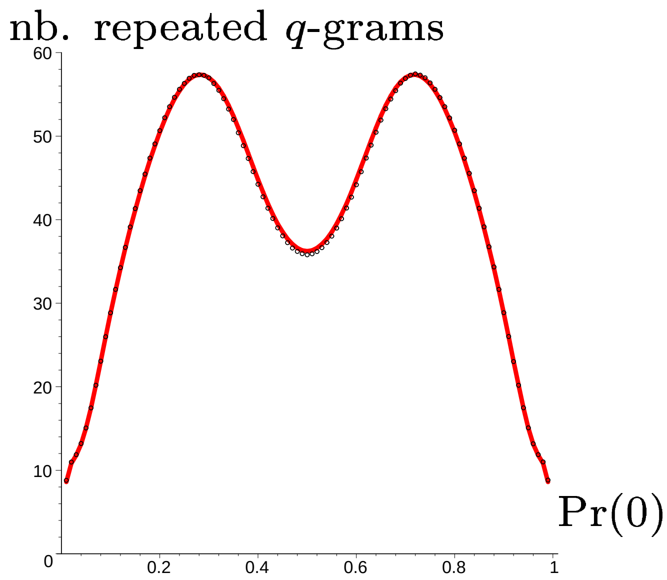
5. mark with letter u all transitions changing a label from 1 to 2 (first repetition)

6. Chomski-Schützenberger algorithm for marked automata

Experimental comparisons - Exp



trie (independent) - urn model



suffix-tree (dependent)

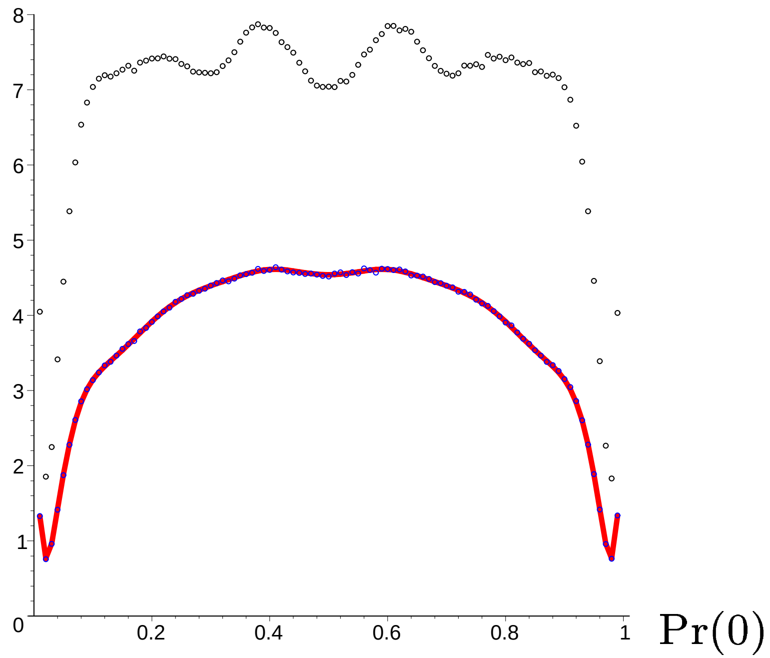
$n = 300 \quad \Sigma = \{0, 1\} \quad s = 2 \quad q = 10$

solid lines: theoretical curve for the trie

dots: simulations

Experimental comparisons - Std. dev.

repeated q-grams



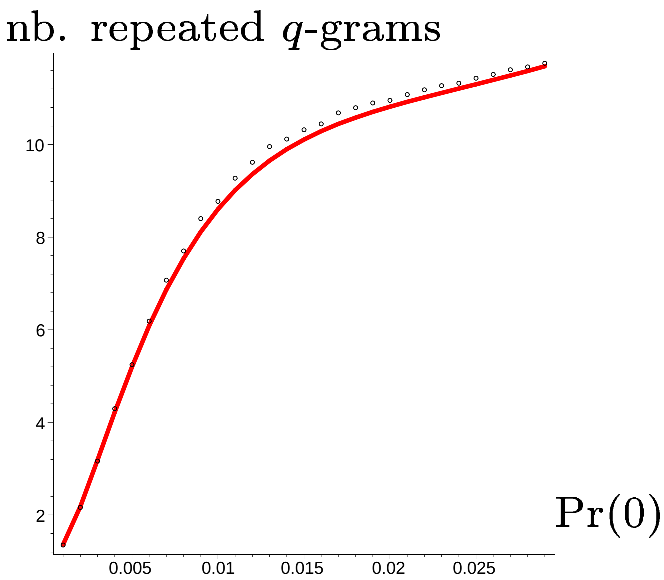
$n = 300 \quad \Sigma = \{0, 1\} \quad s = 2 \quad q = 10$

theoretical σ - trie (solid line)

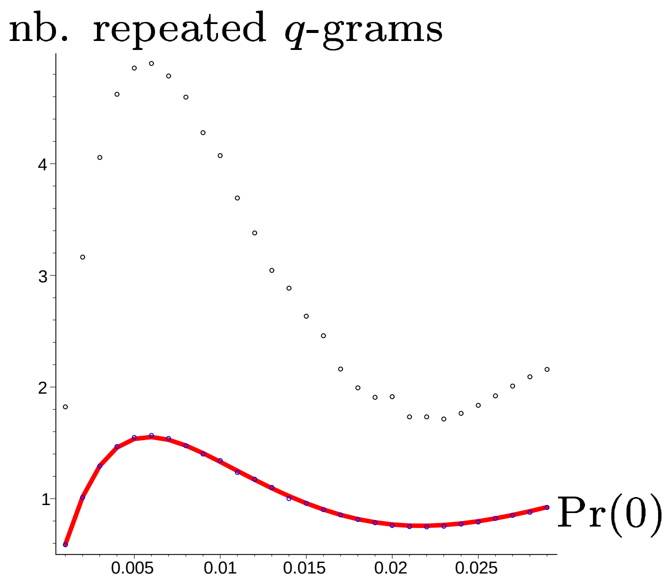
simulations for σ trie (blue circles)

simulations for σ suffix-tree (black circles)

Small p



expectation



standard deviation

$$n = 300 \quad s = 2 \quad q = 10$$

$$(0.995 + 0.005u)^{300} = \quad 0.2223 + 0.3351u$$
$$+ 0.2518u^2 + 0.1257u^3 + 0.047u^4 \ldots$$

Common q-grams to 2 sequences

Equivalent problems

Input: an alphabet Σ ($|\Sigma| = s$), an integer q, 2 random sequence S_1 and S_2 of size $n + q - 1$

Dependent model

1. number of repeated q-grams
2. number of bicolor nodes at depth q when superposing colored suffix-trees build on S_1 and S_2
3. number of intersections of two random walk of length n over the de Bruijn graph $B(s, q)$

Independent model

1. number of repeated q-grams
2. number of bicolor nodes at depth q when superposing two colored tries build eachwith n random keys over Σ
3. number of intersections of two random walks of length n over a complete graph $K(s^q)$
4. number of urns with bicolor collisions in a system of s^q urns in which n black and n white balls are thrown

Previous results

– P. Flajolet, P. Kirschenhofer, and R. F. Tichy - 1988, W. Szpankowski -1993

asymptotically, all words of size $\log(n)/H$ are present in a text of size n

(H Renyi-entropy of the alphabet)

$H = \log \omega_{\min}$ where $\omega_{\min}$ is the minimum of the probability of the letters of the alphabet

Analysis of the urn model

– g.f. and moments

double poissonization-depoissonization

$$F(z,t,u) = \prod_{0\le i\le s^q-1} \left(e^{p_i(z+t)} + (u-1)(e^{p_i z} + e^{p_i t} - 1)\right)$$

z black balls , t white balls $\qquad$ u bicolor collisions

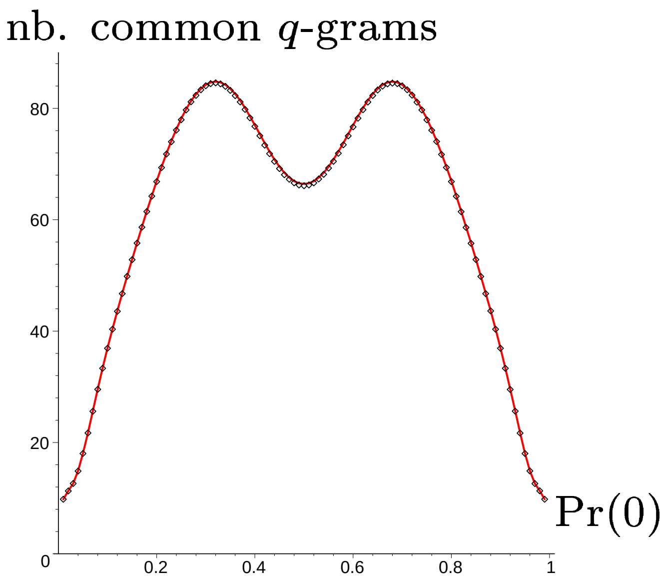
$$\mu_n = m - [z^n t^n] \left.\frac{\partial F(z,t,u)}{\partial u}\right|_{u=1}$$
$$= m - \sum_i \left(e^{-\theta_i}(2-e^{-\theta_i}) - \frac{\theta_i^2 e^{-\theta_i}}{n}(1-e^{-\theta_i})\right) + o(1)$$

$$\sigma_n^2 \approx \sum_i e^{-\theta_i}(2-e^{-\theta_i})\left(1-e^{-\theta_i}(2-e^{-\theta_i})\right)$$
$$-\frac{2}{n}\left(\left(\sum_i \theta_i e^{-\theta_i}\left(1-e^{-\theta_i}\right)\right)^2 - \sum_i \theta_i^2 e^{-2\theta_i}(1-e^{-\theta_i})^2\right)$$

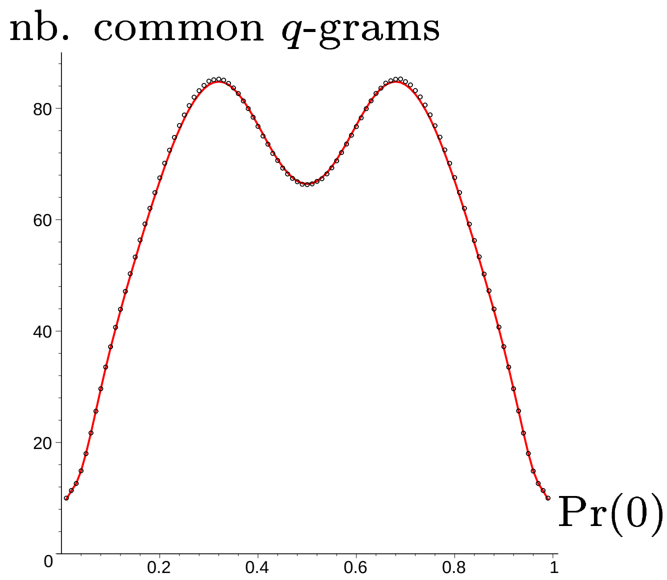
– Poisson convergence

Chen-Stein + coupling (reverse Barbour-Holst)

Experimental comparisons - Exp



trie (independent) - urn model



suffix-tree (dependent)

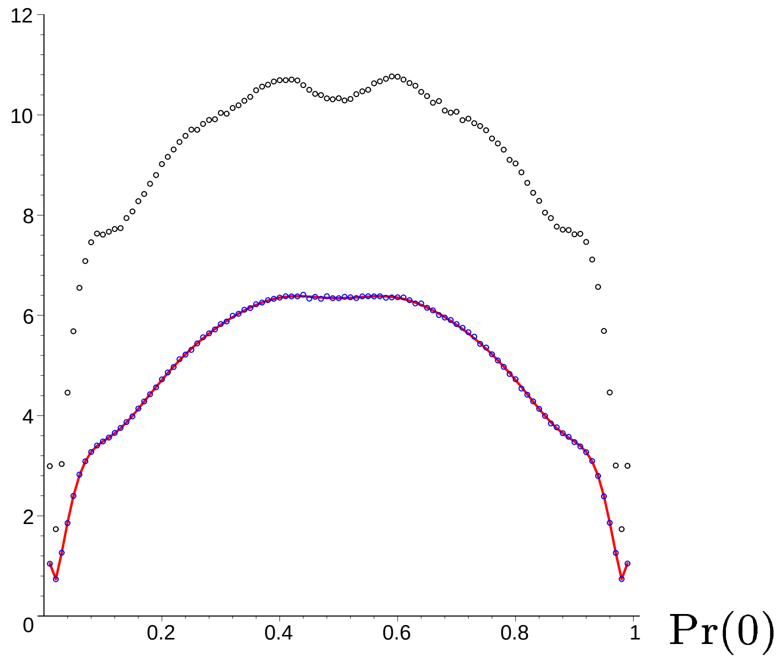
$n = 300 \quad \Sigma = \{0, 1\} \quad s = 2 \quad q = 10$

solid lines: theoretical curve for the trie

dots: simulations

Experimental comparisons - Std. dev.

common q-grams



$n = 300 \quad \Sigma = \{0, 1\} \quad s = 2 \quad q = 10$

theoretical σ - trie (solid line)

simulations for σ trie (blue circles)

simulations for σ suffix-tree (black circles)

Cost of summations

$\Sigma = \{1,2,3,4\}$, $s = |\Sigma|$ $\qquad m = s^q$

group urns by families of urns with equal probability

$|w| = q$, $\quad |w_i| = q_i$ number letters equal to i,

$q = q_1 + q_2 + q_3 + q_4 \qquad$ population of $(q_1, q_2, q_3, q_4) = \dfrac{q!}{q_1! q_2! q_3! q_4!}$

Number of families $C_{q,s}$ (cost of summation)

$$C_{q,s} = \text{compositions with } s \text{ summands} \geq 0 \text{ of } q$$
$$= \text{compositions with } s \text{ summands} > 0 \text{ of } q+s$$

$$C_s(z) = \left(\frac{z}{1-z}\right)^s$$

$$C_{q,s} = [z^{q+s}] \left(\frac{z}{1-z}\right)^s = \binom{q+s-1}{s-1}$$

ADN: $C_{10,4} = 286 \qquad$ Proteins: $C_{3,20} = 1540$

Computing the moments (repeated q-grams)

The values of q_1 to q_{i-1} have been computed previously when Procedure Calcsum is entered and $d = s - i$.
$s = |\Sigma|$ and q are handled as global constants.

Procedure Calcsum (f, d, n, ϕ)**:**

$i = s - d \qquad u = \sum_{k=1}^{i-1} q_k$

If $d > 1$ **Then**

For j **To** $s - u$ **Do**

$q_i = j \qquad f =$ **Calcsum**$(f, d-1, n, \phi)$

End of for

Else

$q_s = q - \sum_{k=1}^{s-1} q_k$

$f = f + \frac{q!}{q_1! q_2! \ldots q_s!} \phi(\theta_{q_1,\ldots,q_s}, n)$

End of if

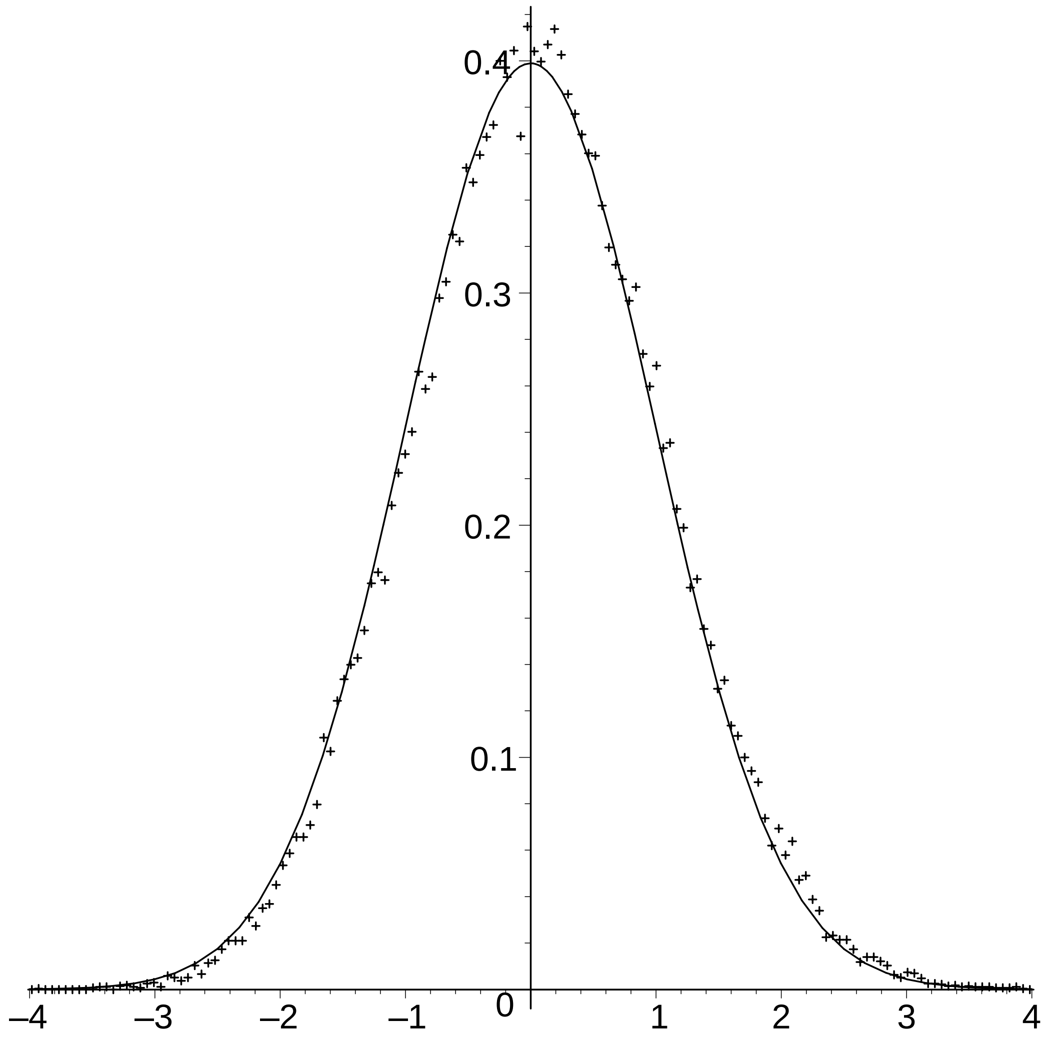
Return (f)

End of procedure

$$\theta_\xi = \theta_{q_1,\ldots q_s} = n \times \omega_1^{q_1} \omega_2^{q_2} \ldots \omega_s^{q_s}$$

$$\phi_1 = \left(e^{-\theta_\xi}(1 + \theta_\xi) + \frac{1}{2n} e^{-\theta_\xi} \theta_\xi^2 (1 - \theta_\xi) \right)$$

$\mu_n = m -$ **Calcsum**$(0, s, n, \phi_1)$

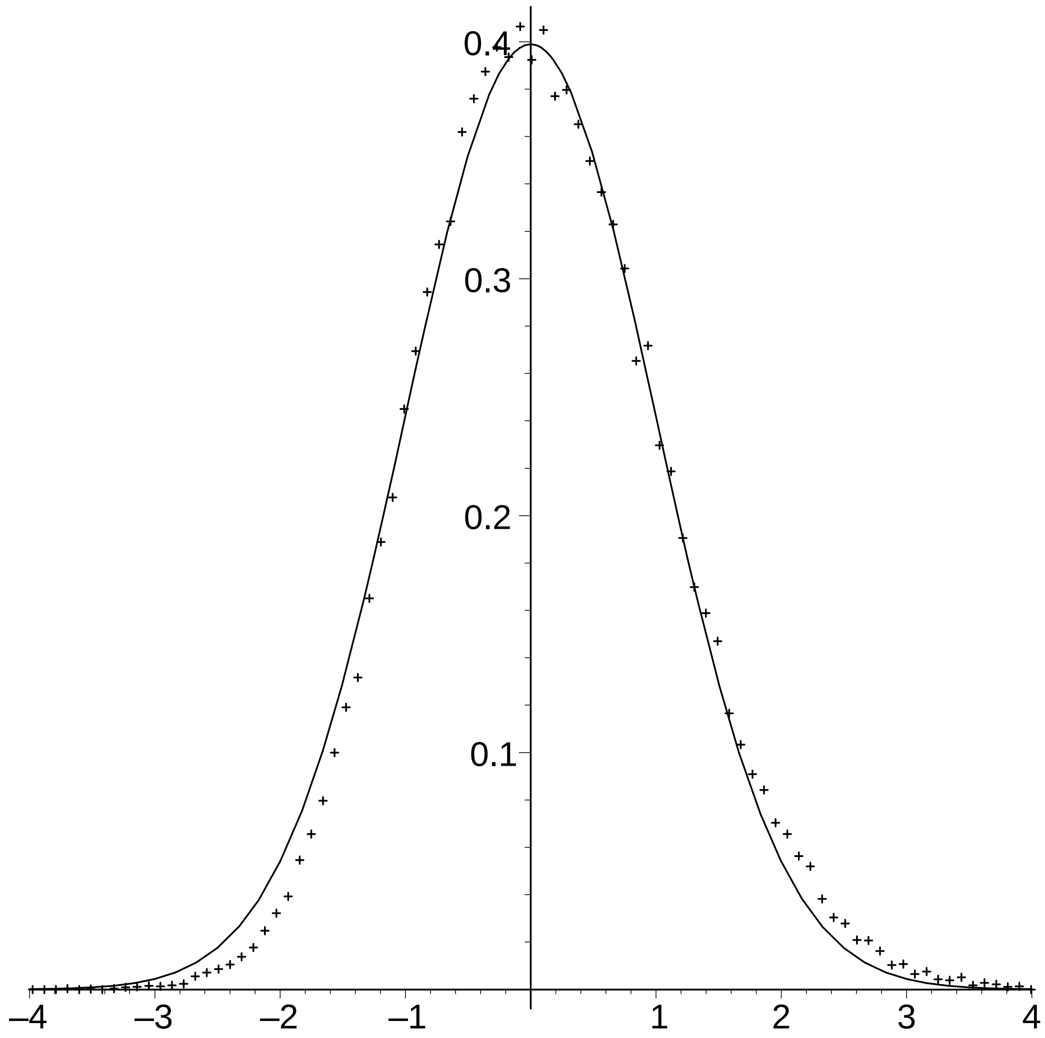


$n_1 = n_2 = 1000$, $q = 12$, $\Sigma = \{0, 1\}$ $p_0 = p_1 = 0.5$

50000 simulations

K number of common 12-grams

Plot of the normalized variable $\widehat{K}$ versus $\mathcal{N}(0, 1)$



$n_1 = n_2 = 1000$, $q = 12$, $\Sigma = \{0, 1\}$ $p_0 = 0.1 \quad p_1 = 0.9$

50000 simulations

K number of common 12-grams

Plot of the normalized variable $\widehat{K}$ versus $\mathcal{N}(0, 1)$

Jokinen-Ukkonen statistics (common q-grams)

urn model

z counts black balls, $p_i = \Pr(\text{black ball falls in urn } i)$

t counts white balls, $x_i = \Pr(\text{white ball falls in urn } i)$

double poissonization-depoissonization, g. f. for one urn: $e^{p_i z} \times e^{x_i t}$

u counts the total number of black balls that are present in urns containing at least one white ball

$$\begin{bmatrix} 1 & (p_i z) & \dots & \frac{(p_i z)^i}{i!} & \dots \\ (x_i t) & u(p_i z)(x_i t) & \dots & \frac{u^i (p_i z)^i}{i!}(x_i t) & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \frac{(x_i t)^j}{j!} & u(p_i z)\frac{(x_i t)^j}{j!} & \dots & \frac{u^i (p_i z)^i}{i!}\frac{(x_i t)^j}{j!} & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}$$

$$F(z,t,u) = \prod_{0 \le i \le s^q - 1} e^{p_i u z + x_i t} - e^{p_i u z} + e^{p_i z} = \sum f_{kab} u^k z^a t^b$$

$f_{kab} = \Pr(k$ black balls in urns with at least 1 white ball

when a white and b black balls are thrown).

Expectation and Standard Deviation

$$p_i = x_i, \quad a = b = n$$

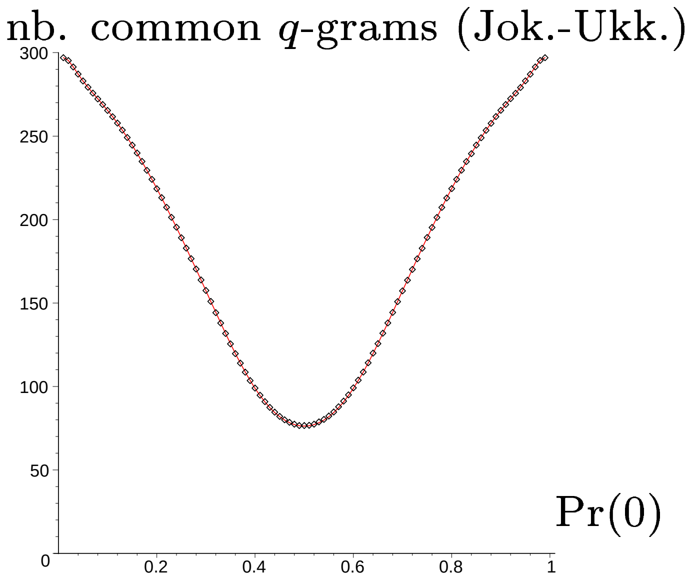
$$n \to \infty, \quad n \times p_i \to \theta_i$$

$$\kappa_i = \sum_i \theta_i \left(1 - e^{-\theta_i}\left(1 - \frac{\theta_i^2}{2n}\right)\right)$$

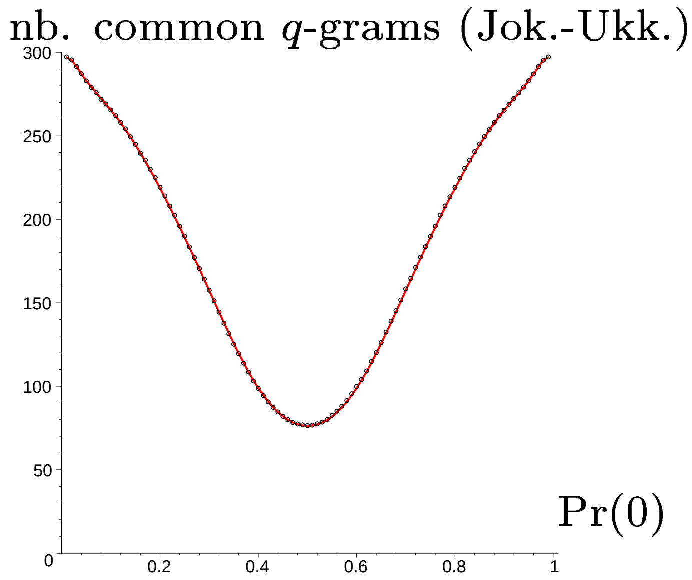
$$\mu_n \approx \sum_i \kappa_i$$

$$\sigma_n^2 \approx \sum_i \kappa_i(\theta_i - \kappa_i) - \frac{1}{n}\left(\left(\sum_i \theta_i \left(1 - e^{-\theta_i}\right)\right)^2 + \left(\sum_i \theta_i^2 e^{-\theta_i}\right)\right)$$

Experimental comparisons - Exp



trie (independent) - urn model



suffix-tree (dependent)

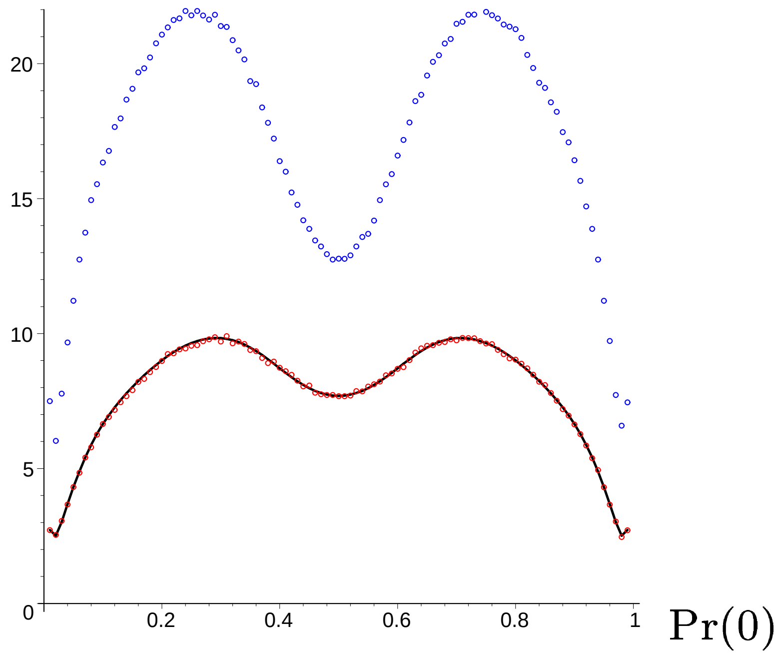
$n = 300 \quad \Sigma = \{0, 1\} \quad s = 2 \quad q = 10$

solid lines: theoretical curve for the trie

dots: simulations

Experimental comparisons - Std. dev.

common q-grams (Jok.-Ukk.)



$n = 300 \quad \Sigma = \{0, 1\} \quad s = 2 \quad q = 10$

theoretical σ - trie (solid line)

simulations for σ trie (blue circles)

simulations for σ suffix-tree (black circles)