Profiles of PATRICIA Tries

Abram Magner and Wojciech Szpankowski
Purdue University
W. Lafayette, IN 47907

June 11, 2015

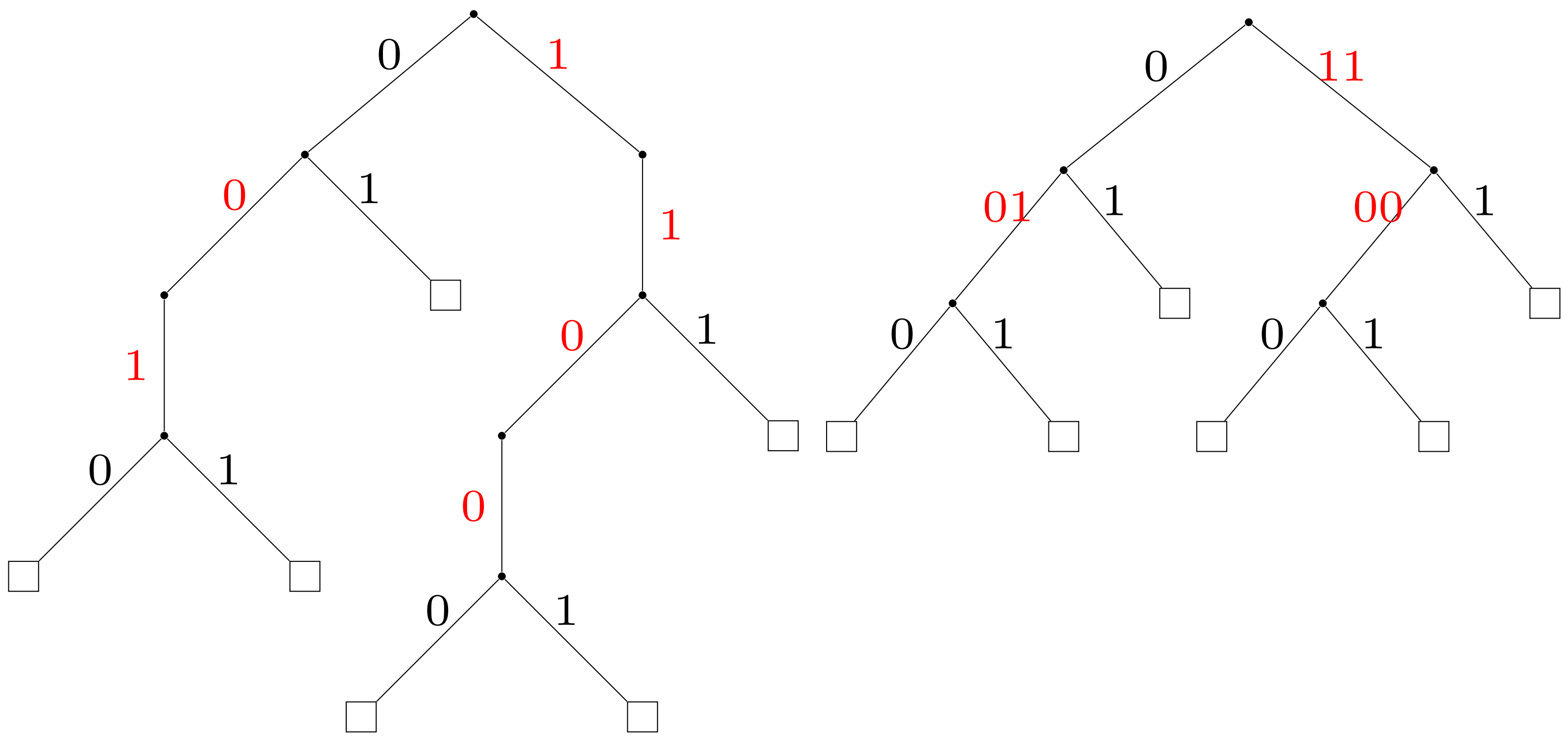
Center for Science of Information
NSF Science & Technology Center

AofA, Strobl, 2015

Outline of the Presentation

1. Tries, PATRICIA Tries, Profiles
2. Expected Value Analysis
   (a) Poisson functional equation
   (b) Mellin transform
   (c) Inverse Mellin transform via saddle point method
3. Limiting Distribution Proof
   (a) Poisson PGF recurrence
   (b) Remainder term via Cauchy integral formula
4. Next Steps: Height

Tries, PATRICIA Tries, Profiles



External profile for the trie:

$B_{6,0} = 0$, $B_{6,1} = 0$, $B_{6,2} = 1$, $B_{6,3} = 1$, $B_{6,4} = 2$, $B_{6,5} = 2$.

External profile for the PATRICIA trie:

$B_{6,0} = 0$, $B_{6,1} = 0$, $B_{6,2} = 2$, $B_{6,3} = 4$.

Profiles

Definition 1 (Internal/External profile). *The internal, external profile at level k of a trie on n strings, $I_{n,k}$, $B_{n,k}$, is the number of internal, external nodes, respectively, at level k.*

Several parameters are of interest in the analysis of algorithms which use digital trees:

- typical depth D_n: the depth of a randomly chosen leaf. Typical search time.
- height H_n: the maximum depth of any leaf. Worst-case search time.
- fillup level F_n: the maximum full level of the trie (i.e., all possible internal nodes exist). Plays a role in analysis of level-compressed tries.

- $\Pr[D_n = k] = \frac{\mathbf{E}[B_{n,k}]}{n}$.
- $H_n = \max\{k | B_{n,k} \neq 0\}$.
- $F_n = \max\{k | I_{n,k} = 2^k\}$.

Memoryless source model: strings are i.i.d., each a sequence of i.i.d. Bernoulli random variables with fixed bias $p > 1/2$.

Prior Work

Trie profiles: Park (2006), and Park, Hwang, Nicodeme, W.S., (2008).

Digital search tree profiles: Expected value fully analyzed by Drmota & W.S. (2011). Variance for the asymmetric case analyzed by Kazemi & Vahidi-Asl (2011).
PATRICIA and DSTs have the same range of polynomial growth, where

$$\mu_{n,k} = \Theta\left(\frac{n^{\beta(\alpha)}}{\sqrt{\log n}}\right), \quad k = \alpha \log n.$$

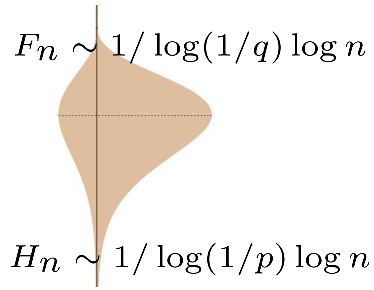
PATRICIA:

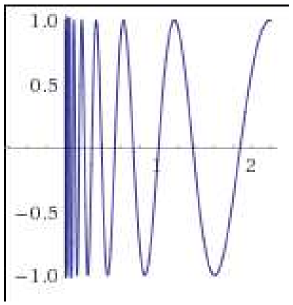
- Knuth (1968) and W.S.(1990): Depth in PATRICIA.
- Pittel & Rubin (1987 and Devroye (1992): Derivation of first two terms of the typical height for symmetric PATRICIA trie:

$$\frac{H_n - \log_2 n}{\sqrt{2 \log_2 n}} \xrightarrow{n \to \infty} 1.$$

- Devroye (2004): showed concentration for *total profile* ($I_{n,k} + B_{n,k}$) under weak probabilistic assumptions.
- Knessl & W.S. (2002): Limiting distribution of height is concentrated on a few points (using WKB method).

Main Results





Asymptotic expansions for mean $\mathbf{E}[B_{n,k}] = \mu_{n,k}$, and variance in the range of polynomial growth (i.e.,

$$k \sim \alpha \log n, \alpha \in (1/\log(1/q), 1/\log(1/p))).$$

For the expected external profile,

$$\mathbf{E}[B_{n,k}] \sim H(\rho, \log n) n^{-\rho+\alpha\log(p^{-\alpha}+q^{-\alpha})} \sim H(\rho) n^{\beta(\alpha)}.$$

Variance:

$$\mathbf{Var}[B_{n,k}] \sim K(\rho, \log n) n^{-\rho+\alpha\log(p^{-\alpha}+q^{-\alpha})} = \Theta(\mathbf{E}[B_{n,k}]).$$

where ρ is a saddle point, and $H(\rho)$ and $K(\rho)$ are fluctuating functions.

- Central limit theorem for profile in range of polynomial growth:

$$\mathbf{E}\left[\exp\left(\tau \frac{B_{n,k} - \tilde{G}_k(n)}{\sigma_{n,k}}\right)\right] = \exp\left(\frac{\tau^2}{2}(1 + O(V_{n,k}^{-1/2}))\right)$$

where $\mu_{n,k} \sim \tilde{G}_k(n)$ and $G_k(z) = \sum_{n=0}^{\infty} \mu_{n,k} \frac{z^n}{n!}$.

Expected Value Analysis

Approach:

- Unlike in tries and DST, path compression means no closed form solution to the relevant recurrence equation.
- To solve for $\mu_{n,k}$ in the interesting range, need information about it in other ranges.

$$\boxed{\text{Poissonization}} \to \boxed{\text{Mellin}} \to \boxed{\text{Inverse Mellin (SP)}} \to \boxed{\text{De-Poissonization}}$$

- Poissonization: Poisson splitting property makes deriving a functional equation easy. Asymptotics of $\mu_{n,k}$ reduced to asymptotics of $\tilde{G}_k(z)$ as $z, k \to \infty$.
- Mellin transform: functional equation is transformed into an algebraic equation.
- Mellin inversion: via saddle point method for asymptotically evaluating integrals.
- De-Poissonization: justify the equivalence $\mu_{n,k} \sim \tilde{G}_k(n)$ as $n \to \infty$.

Expected Value Derivation

Poisson functional equation: Define $G_k(z) = \sum_{n=0}^{\infty} \mu_{n,k} \frac{z^n}{n!}$ and $\tilde{G}_k(z) = e^{-z} G_k(z)$. Then

$$\tilde{G}_j(z) = \tilde{G}_{j-1}(pz) + \tilde{G}_{j-1}(qz) + \tilde{W}_j(z),$$

$$\tilde{W}_j(z) = e^{-pz}[\tilde{G}_j - \tilde{G}_{j-1}](qz) + e^{-qz}[\tilde{G}_j - \tilde{G}_{j-1}](pz).$$

Trie functional equation:

$$\tilde{T}_j(z) = \tilde{T}_{j-1}(pz) + \tilde{T}_{j-1}(qz).$$

where $\tilde{T}_j(z)$ is the Poisson transform of the profile.

DST functional equation:

$$\Delta'_{k+1}(z) + \Delta_{k+1}(z) = \Delta_k(pz) + \Delta_k(qz),$$

where $\Delta_k(z)$ is the Poisson transform of the profile in DST.

PATRICIA Solution

Mellin transform: We get an exact (implicit) formula for $G_k^*(s)$ by unraveling the recurrence:

$$G_k^*(s) = (p^{-s} + q^{-s}) G^*{}_{k-1}(s) + W_k^*(s) = T(s)^k \Gamma(s+1) A_k(s),$$

with

$$T(s) = p^{-s} + q^{-s},$$

$$A_k(s) = 1 + \sum_{j=1}^{k} T(s)^{-j} \sum_{m=j}^{\infty} T(-m)(\mu_{m,j} - \mu_{m,j-1}) \frac{\Gamma(m+s)}{\Gamma(s+1)\Gamma(m+1)}$$

Compare with tries: $\mathcal{T}_j^*(s) = T(s)^k \mathcal{T}_0^*(s)$ where $\mathcal{T}_0^*(s) = \Gamma(s+1) g(s)$ for nice $g(s)$.

Properties of $A_k(s)$:
The function $A_k(s)$ is entire, with zeros at $-k, -k+1, \ldots, -1$ which cancel out poles of $\Gamma(s+1)$. Also,

$$\lim_{k \to \infty} A_k(s) = A(s)$$

pointwise for all s.

The fundamental strip of $\tilde{G}_k(z)$ is $\Re(s) \in (-k-1, \infty)$.

Inverse Mellin: Saddle Point Method

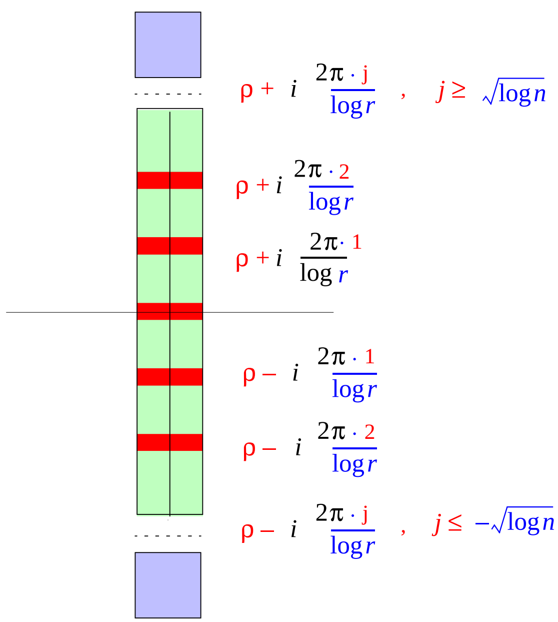
By depoisonization we have $\tilde{G}_k(n) \sim \tilde{G}_k(z)$, where recall

$$\begin{aligned}\tilde{G}_k(n) &= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} A_k(s)\Gamma(s+1)n^{-s}(p^{-s}+q^{-s})^k ds \\ &= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} A_k(s)\Gamma(s+1)\exp(h(s)\log n)ds, \qquad k = \alpha \log n.\end{aligned}$$

The saddle point equation $h'(s) = 0$ has a unique real root:

$$\rho = \frac{-1}{\log r}\log\left(\frac{\alpha \log q^{-1} - 1}{1 - \alpha \log p^{-1}}\right), \qquad \frac{1}{\log q^{-1}} < \alpha < \frac{1}{\log p^{-1}}.$$

There are **infinitely many** saddle points $\rho + it_j$ for $t_j = 2\pi j/\log r$, $j \in \mathbb{Z}$.



Some Comments:

1. $\rho \to \infty$ as $\alpha \downarrow 1/\log q^{-1} = \alpha_1$.

2 $\rho \to -\infty$ when $\alpha \uparrow 1/\log p^{-1}$.

3. In tries the saddle points coalesce with poles of the $\Gamma(s+1)$ function at $s = -2, -3, \ldots$.
In PATRICIA these poles are cancel out.

Limiting Distribution Proof

Poisson PGF recurrence:
Define $Q_k(u,z) = \sum_{n=0}^{\infty} \mathbf{E}[u^{B_{n,k}}]\frac{z^n}{n!}$ and $\tilde{G}_k(u,z) = e^{-z}G_k(u,z)$. Then

$$\tilde{Q}_k(u,z) = \tilde{Q}_{k-1}(u,pz) + \tilde{Q}_{k-1}(u,qz) + e^{-pz}(\tilde{Q}_k - \tilde{Q}_{k-1})(u,qz) \\ + e^{-qz}(\tilde{Q}_k - \tilde{Q}_{k-1})(u,pz).$$

Define $\tilde{l}_k(w,x) = \log(\tilde{Q}_k(w,x))$. With $u = e^{it/\sigma_{n,k}}$,

$$\tilde{l}_k(u,z) = \tilde{G}_k(z)\frac{\tau}{\sigma_{n,k}} + \tilde{V}_k(z)\frac{\tau^2}{\sigma_{n,k}^2} + O(\sigma_{n,k}^{-1}) + \frac{\tau^3}{\sigma_{n,k}^3}R[\ell]_k(u,z).$$

Goal:
Show that $R[\ell]_k(u,z)$ is negligible with respect to the other terms which are of order $O(n^{\beta(\alpha)})$, where $\beta(\alpha)$ is the polynomial order of growth of $\mu_{n,k}$ and $V_{n,k}$.

Limiting Distribution Proof

Remainder term via Cauchy integral formula:

$$\frac{1}{3!}R[\tilde{l}]_k(u,z) = R_{1,k}(u,z) + R_{2,k}(u,z),$$

where

$$R_{1,k}(u,z) = \sum_{j=0}^{k} \binom{k}{j} \frac{1}{2\pi i} \oint_{\mathcal{C}} \frac{p^j q^{k-j} z + \log(1 + (w-1)p^j q^{k-j} z e^{-p^j q^{k-j} z})}{(w-1)^3 (w-u)} \, \mathrm{d}w$$

$$R_{2,k}(u,z) = \sum_{j=0}^{k} \sum_{m=0}^{k-j} \binom{k-j}{m} \frac{1}{2\pi i} \oint_{\mathcal{C}} \frac{\tilde{h}_j(w, p^m q^{k-j-m} z)}{(w-1)^3 (w-u)} \, \mathrm{d}w,$$

where

$$\tilde{h}_j(w,x) = \log\left(1 + \frac{(Q_j - Q_{j-1})(w,px) + (Q_j - Q_{j-1})(w,qx)}{Q_{j-1}(w,px) Q_{j-1}(w,qx)}\right).$$

Here, $\mathcal{C}$ is a circle centered at 1 and containing u in its interior. Both $R_{1,k}(u,z)$ and $R_{2,k}(u,z)$ are $O(n^{\beta(\alpha)})$ for $|z| = n$ in a cone.

Limiting Distribution Proof

Bounding $R_{1,k}(u, z)$: Laurent expansion of the integrand gives

$$R_{1,k}(u, z) \sim \sum_{j=0}^{k} \binom{k}{j} (p^j q^{k-j} z)^3 e^{-3p^j q^{k-j} z}.$$

Range: $j = o(\log n)$ or $j \sim k$.
Contribute negligibly, because

$$\binom{k}{j} \le k^j / j!, \quad \binom{k}{j} \le k^{k-j}/(k-j)!,$$

and the other factors are bounded, so these terms are at most $e^{o(\log n)} = n^{o(1)}$.

Range $j, k - j = \Theta(\log n)$.
Write each term as $e^{g(j)}$ using Stirling's formula, and, taking derivatives of $g(j)$ to find the largest term, we find that the sum is at most $\tilde{\Theta}(n^{\beta(\alpha)})$.

Bounding $R_{2,k}(u, z)$: More difficult.
Requires precise lower and upper bounds on $(Q_j - Q_{j-1})(w, x)$, which rely on knowledge of $\mu_{m,j}$ for $m = \Theta(j)$ (so to the right of the saddle point range). See the paper.

Next Steps: Height

It is known (Pittel, Devroye) that for symmetric PATRICIA, the height H_n behaves like:

$$H_n = \log_2 n + \sqrt{2 \log_2 n} + O(1) \text{ whp}$$

Next Steps: Height

It is known (Pittel, Devroye) that for symmetric PATRICIA, the height H_n behaves like:

$$H_n = \log_2 n + \sqrt{2\log_2 n} + O(1) \quad \text{whp}$$

Conjecture 1. *For asymmetric PATRICIA we claim that ($p \neq q = 1 - p$)*

$$H_n = \log_{1/p} n + \log_{p/q} \log n + O(\log\log\log) \quad \textit{whp.}$$

Need more precise estimates of $\mu_{n,k}$ and $V_{n,k}$ to the right of the saddle point interval.

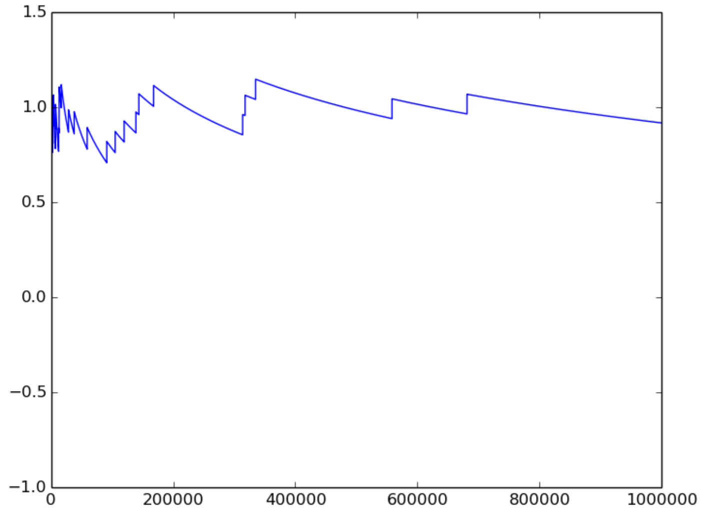
Lower bound:

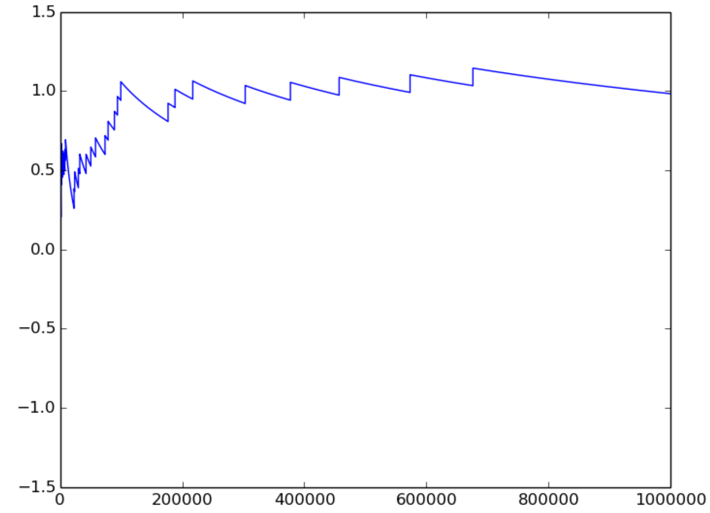
$$\Pr[H_n < k] \leq \Pr[B_{n,k} = 0] \leq \frac{\mathbf{Var}[B_{n,k}]}{\mathbf{E}[B_{n,k}]^2}$$

Upper bound:

$$\Pr[H_n > k] \leq \sum_{j>k} \mathbf{E}[B_{n,j}].$$

Experimental Results for the Height





New Book on Pattern Matching

How do you distinguish a cat from a dog by their DNA? Did Shakespeare really write all of his plays?

Pattern matching techniques can offer answers to these questions and to many others, from molecular biology, to telecommunications, to classifying Twitter content.

This book for researchers and graduate students demonstrates the probabilistic approach to pattern matching, which predicts the performance of pattern matching algorithms with very high precision using analytic combinatorics and analytic information theory. Part I compiles known results of pattern matching problems via analytic methods. Part II focuses on applications to various data structures on words, such as digital trees, suffix trees, string complexity and string-based data compression. The authors use results and techniques from Part I and also introduce new methodology such as the Mellin transform and analytic depoissonization.

More than 100 end-of-chapter problems help the reader to make the link between theory and practice.

Philippe Jacquet is a research director at INRIA, a major public research lab in Computer Science in France. He has been a major contributor to the Internet OLSR protocol for mobile networks. His research interests involve information theory, probability theory, quantum telecommunication, protocol design, performance evaluation and optimization, and the analysis of algorithms. Since 2012 he has been with Alcatel-Lucent Bell Labs as head of the department of Mathematics of Dynamic Networks and Information. Jacquet is a member of the prestigious French Corps des Mines, known for excellence in French industry, with the rank of "Ingenieur General". He is also a member of ACM and IEEE.

Wojciech Szpankowski is Saul Rosen Professor of Computer Science and (by courtesy) Electrical and Computer Engineering at Purdue University, where he teaches and conducts research in analysis of algorithms, information theory, bioinformatics, analytic combinatorics, random structures, and stability problems of distributed systems. In 2008 he launched the interdisciplinary Institute for Science of Information, and in 2010 he became the Director of the newly established NSF Science and Technology Center for Science of Information. Szpankowski is a Fellow of IEEE and an Erskine Fellow. He received the Humboldt Research Award in 2010.

Cover design: Andrew Ward

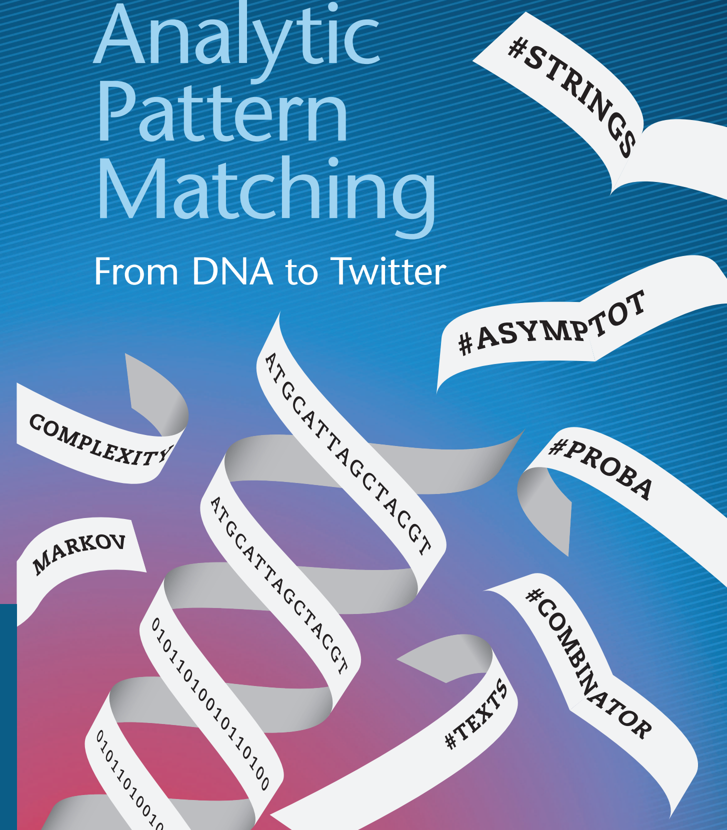
CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org
ISBN 978-0-521-87608-7
9 780521 876087

Jacquet and Szpankowski Analytic Pattern Matching CAMBRIDGE

Philippe Jacquet and
Wojciech Szpankowski

Analytic Pattern Matching

From DNA to Twitter



Book Contents

That's It

THANK YOU