The Digital Tree Process

Julien Clément\textsuperscript{1} \quad Mark Daniel Ward\textsuperscript{2}, *

CNRS, research team Algorithms of GREYC\textsuperscript{1}

Department of Statistics, Purdue University\textsuperscript{2}

December 16, 2011

\textit{Dedicated to the memory of Philippe Flajolet.}

*Supported by NSF Science & Technology Center grant CCF-0939370.
The analyses of digital tree processes pervade Philippe Flajolet's work.

Based on the thumb rule in a dictionary Dynamical structure

Principles attributed to Thue (1912) by Knuth. (1960) Discovered by De la Briandais (1959) et Fredkin

trie ≡ tree + retrieval.
Two points of view

Tries

- Tries as a *data structure*
- Tries as a *partitioning digital process*

The digital tree process gave rise to many algorithmic variants:

1. PATRICIA trees,
2. digital search trees
3. LC-tries
4. hybrid trie structures (e.g., array-trie, bst-trie, list-trie)

E.g., *Digital Search Trees Revisited*, PF and R. Sedgewick (1986); and *The Analysis of Hybrid Trie Structures* (1998) and *Dynamical Sources in Information Theory* (2001), both by J. Clément, PF, and B. Vallée.
Tries are **dynamic data structures** that store **randomly generated words**. “Dynamic” because tries **grow** as more words are inserted.

First paragraph of *Moby Dick* (H. Melville)

---

*figure © R. Sedgewick*

**Parameters:** 
- size (memory usage)
- external path length (searching)
- height
Motivations

Conceptually, the digital tree process can appear at unexpected places, and Philippe liked that a lot.

- recursive definition $\rightarrow$ functional equations via generating functions.
- database management; data mining
- data compression; closely related to Lempel-Ziv schemes
- efficient communication protocols; conflict resolution
- leader election and connections to distributed computing
- probabilistic counting
- hashing; bucket sorting (e.g., $\geq 1$ string per leaf)
- polynomial factorization
- dictionary
- sorting and searching [Knuth 73]
- set intersection, set union [Trabb Pardo 78]
- multiway branching for generalized (non-binary) alphabet
Recursive Definition


(Uncompressed!) Trie for sequences

\[
\begin{align*}
A &= 111 \ldots & D &= 101 \ldots \\
B &= 011 \ldots & E &= 1100 \ldots \\
C &= 1101 \ldots & F &= 010 \ldots
\end{align*}
\]

The splitting groups are “the ‘heads’ group and the ‘tails’ group.”

Much later, for instance, in *The Ubiquitous Digital Tree* (STACS, 2006), he defines a trie recursively. For a set of strings \( \omega \),

\[
\text{trie}(\omega) := \begin{cases} 
\emptyset & \text{if } \omega = \emptyset, \\
\sigma & \text{if } \omega = \{\sigma\}, \\
(\bullet, \text{trie}(\omega \setminus 0), \text{trie}(\omega \setminus 1)) & \text{otherwise}
\end{cases}
\] (1)

“The motto here is thus simply ‘filter and shift left’.”
Recursive Description of Tries

From Philippe’s “Saga of Digital Trees”, part of the Colloquium for Jacques Morgenstern in 2003:

Compare-exchange based on successive \textit{bits} of data.
place 0’s on left, 1’s on right;
recurse.

The trie splitting process (Fredkin)
One of Philippe’s tries (TikZ rendering, used in several of his talks) built on 500 uniform binary sequences, with 741 internal nodes, and height 17:
Probability Models

The probability models for tries assume independence between the strings stored at the leaves.

The characters within the string may be independent, uniform (i.e., $p = q = 1/2$), independent, biased ($p \neq q$), or have Markov dependence, or even have a dynamical source.

Philippe’s work also laid a foundation for analysis of suffix trees, in which the strings are dependent: they are suffixes of a common string. [E.g., Nicodème; Jacquet, Szpankowski; J. Fayolle, MDW.]

First-order, expected behavior of suffix tree parameters often agrees with analogous behavior for tries built over independently-generated strings.

First-order variance, higher moments, and also second-order terms of expectation, are often different in suffix trees vs tries built over independent strings.
Comparison with Suffix Trees

\[ X_1, X_2, X_3, \ldots = 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, \ldots \]

Suffixes of the sequence:

\[ S_1 = 0, 1, 0, 1, 1, 0, \ldots \]
\[ S_2 = 1, 0, 1, 1, 0, 0, \ldots \]
\[ S_3 = 0, 1, 1, 0, 0, 1, \ldots \]
\[ S_4 = 1, 1, 0, 0, 1, 1, \ldots \]
\[ S_5 = 1, 0, 0, 1, 1, 1, \ldots \]
\[ S_6 = 0, 0, 1, 1, 1, 1, \ldots \]
\[ S_7 = 0, 1, 1, 1, 1, 0, \ldots \]
\[ S_8 = 1, 1, 1, 1, 0, 0, \ldots \]
\[ S_9 = 1, 1, 1, 0, 0, 0, \ldots \]
\[ S_{10} = 1, 1, 0, 0, 0, 0, \ldots \]
\[ S_{11} = 1, 0, 0, 0, 0, 0, \ldots \]
\[ S_{12} = 0, 0, 0, 0, 0, 1, \ldots \]
We see Philippe building on algebraic properties, e.g., set constructions, multisets, translations to generating functions, and complex-valued analysis [inheriting from De Bruijn, Knuth, Henrici, others].

“The power [of these results] comes from the fact that most parameters of interest on trees are definable as additive-multiplicative combinations of similar or simpler parameters on subtrees, so that a large number of equations can be written systematically.” [PF, D. Sotteau 1982; emphasis added]

Philippe uses generating functions to unify earlier analyses of

- the family of sets whose associated tree has height \( \leq k \)
- generalized versions for leaves with \( \leq b \) nodes (buckets)
- total number of nodes in tree when height is \( \leq k \)
- path length
In 1982, Philippe was already synthesizing connections among different analysis of tries

- **Collision resolution in networks** [G. Fayolle, PF, M. Hofri 1982]
  - Time to resolve $n$ collisions in an open stack protocol network:
    \[ \alpha_n = An + n\phi(n) + O(n / \log n) \]

- **Digital sorting and searching** [Knuth 1973]
- **Dynamic Hashing** [PF, J. M. Steyaert 1982]
- **Polynomial Factorization** [PF, J. M. Steyaert 1982]

The function $\phi(n)$ is a fluctuating function with small amplitude.
Some Trie Parameters

Commonalities: Trie arise in unexpected situations. The analysis often concerns asymptotic properties of a trie parameter (often called valuations by Philippe), e.g.,

- **path length**: sum (over all leaves) of distances from the root to the leaves,
- **total number of nodes**: also called the size of the tree,
- **height**: maximum distance from root to a leaf,
- **number of unary nodes (leaves),**

Philippe was a master at synthesizing ideas and making generalizations of results, especially results that were previously only known in special cases or certain situations.
Basic recurrence (binary case)

Unbiased memoryless source $p = 1/2$.

This decomposition gives for path length

$$L_n = n + \sum_{k=0}^{n} \frac{(n)}{2^n} (L_k + L_{n-k})$$

So that:

$$\hat{L}(z) = \sum_{n \geq 0} L_n \frac{z^n}{n!} = z(e^z - 1) + 2e^{z/2}\hat{L}(z/2).$$

Iterating we obtain

$$\hat{L}(z) = \sum_{k \geq 0} z(e^z - e^{(1-\frac{1}{2k})z}).$$

N.B. An harmonic sum ! (see talk by P. Dumas)
Fundamental generalization:

For recursively defined parameters, the translation schemes (17), (18) lead to functional difference equations of the form

\[ \phi(x) = a(x) \phi \left( \frac{x}{2} \right) + b(x). \]  

(19)

The \( a, b \) are toll functions that depend on the trie parameter.

These can normally be solved by iteration, so that

\[ \phi(x) = \sum_{k \geq 0} b(x2^{-k}) \cap \sum_{j=0}^{k-1} a(x2^{-j}). \]

In the frequent case \( a(x) = A e^{cx} \), (20) further simplifies to

\[ \phi(x) = \sum_{k \geq 0} A^k b(x2^{-k}) \exp(c(1-2^{-k})x). \]  

(20)
For the path length $S_n$:

$$S(x) = \sum (e^{-x\log(1-2^{-k})^{-1} - 1})$$

Such periodicities are of frequent occurrence in the analysis of algorithms and they have here a clear origin in regularly spaced singularities of functions of the type $(1-2^s)^{-1}$. An alternative derivation, which avoids the exponential approximation is based on the observation that $S_n$ is itself a harmonic sum:

Philippe was using saddle points, Mellin transform, and making precise characterization of periodicities, in 1982, more than 10 years before the first Analysis of Algorithms meeting.
Figure 11. Possible ways to obtain the asymptotic mean value in the Bernoulli model from the exact mean value in the Poisson model.
The precise analysis of trie parameters demands careful handling of fluctuations. E.g., the expected size (number of nodes) in a trie built over uniform binary sequences is

\[
\frac{n}{\log 2}(1 + \epsilon(\log n)) + o(n),
\]

where \(\epsilon(x)\) is again a fluctuating function with amplitude less than \(10^{-5}\).
In the three decades since the early 1982 paper, Philippe became a master of the analysis of generating functions, and he went on to characterize many trie parameters under more complicated probabilistic assumptions.

We briefly outline the way that the analysis changes, depending on the branching probabilities $p_1, p_2, \ldots, p_m$ related to the $m$ letters in the alphabet (e.g., $m = 2$ in the binary case).

When the trie is built over uniform binary sequences, the oscillating term arises from the appearance of complex poles, for instance, at

$$z_k = -1 + \frac{2ik\pi}{\log 2}, \quad k \in \mathbb{Z}.$$
Oscillations also arise in the asymptotic growth of parameters of tries built over non-uniform tries (i.e., biased probability model).

E.g., the expected size of a trie with branching probabilities $p, q$ such that $\frac{\log p}{\log q} \in \mathbb{Q}$ is

$$\sim \frac{n}{H}(1 + \epsilon(\ln n)),$$

where $\epsilon$ is oscillating with small amplitude, and

$$H = p \log (1/p) + q \log (1/q)$$

is the entropy of the source that generates the strings.

When $\frac{\log p}{\log q} \notin \mathbb{Q}$, the expected size is $\sim n/H$, with no oscillations in the leading term.
In $b$-tries, there are $\leq b$ strings per leaf (rather than one string per leaf).

The probability that a $b$-trie on $n$ strings has height $\leq h$ is

$$\exp \left( -\frac{2^b u(n) 2^{-b\delta}}{(b + 1)!} \right) (1 + O((\log n)^{b+1}/n^{1/b})), $$

where $u(n)$ is the fractional part of $(1 + 1/b) \log_2 n$, and $\delta = h - \lfloor(1 + 1/b) \log_2 n\rfloor$, and $h$ is in a central region around $(1 + 1/b) \log_2 n$.

The average height is $\bar{H}_n = (1 + 1/b) \log_2 n + P((1 + 1/b) \log_2 n) + o(1)$ where $P$ is periodic.

The average size is $\bar{S}_n = Q((1 + 1/b) \log_2 n)n^{1+1/b}(1 + O(1/(\log n)^{b-1})), $ where $Q$ is periodic.
Timeline: Asymptotic Analysis of Tries [adapted, PF 2010]

- 1965, De Bruijn, Knuth analyze tries built over uniform strings, \( p = q = 1/2 \); oscillations exhibited
- 1973: Knuth (TAOCP, vol. 2) discusses biased case
- 1982: PF makes connections among several types of related analysis; towards systematic treatments
- 1986: Fayolle, PF, Hofri study periodicity criterion, see also Schachinger [2000], and Jacquet, Szpankowski, Tang [2001]
- 2010: PF, Roux, Vallée, convergence to asymptotic regime is very slow and depends on fine arithmetic properties of probabilistic model.
Expected Size of Trie: \( m \)-ary Case

- If \( \frac{\log p_j}{\log p_k} \) are rational for all \( 1 \leq j, k \leq m \), expected size has first-order periodicities:
  \[
  \overline{S}_n = \frac{n}{H} + n\phi(\log n) + O(n^{1-A}), \text{ for } A > 0,
  \]
  where \( H = \sum p_j \log (1/p_j) \) is the entropy of the source.

- If at least one \( \frac{\log p_j}{\log p_k} \) is irrational, then expected size is:
  \[
  \overline{S}_n = \frac{n}{H} + O(n \exp(-\theta \sqrt{\log n})), \text{ for } \theta > 1.
  \]
  “This is better than \( n/(\log n)^a \), any \( a \); much worse than \( n^{1-\epsilon} \), any \( \epsilon \)” (“no oscillation, but poor error term”)

- “For remaining ‘Liouvillean sources’ (rare), error term can come arbitrarily close to \( o(n) \).”

Of course, the set \( \{(p_1, \ldots, p_m) \mid \text{at least one of } \frac{\log p_j}{\log p_k} \text{ is irrational}\} \) has measure 1 in the \( m \)-ary space of all \( m \)-tuples. So this is actually the general situation.
Singularity Analysis

The geometry of the poles

\[
\frac{1}{1 - p_1^s - \cdots - p_m^s}
\]

plays a central role in asymptotic analysis. See especially Digital Trees and Memoryless Sources: from Arithmetics to Analysis, PF, M. Roux, B. Vallée, 2010.

The importance of the geometry during inverse Mellin analysis. Need integration path avoiding poles and estimates of global contributions.
Philip loved to check theoretical, asymptotic results using empirical sources or simulations. E.g., *The Analysis of Hybrid Tree Structures* by J. Clément, P. Flajolet, B. Vallée [SODA, 1998]

*Moby Dick data.* The evolution of insertion costs [left: array-trie, middle: list-trie, right: bst-trie], or equivalently negative search costs, shows an unclear tendency to increase as the number of data items $n$ increases, and there is a fairly large variability of numerical data,

The presentation obtained by plotting against $\log n$ the costs averaged over successive batches of 10 insertions exhibits more clearly the logarithmic trends,

and leads to empirical formulæ for the search costs in array-tries, list-tries, and bst-tries:

$$E_n[R^*] \approx 0.8 \log n, \quad E_n[R^*] \approx 3.0 \log n, \quad E_n[R] \approx 1.0 \log n.$$
Philippe, you are greatly missed.

Your friendship, guidance, and leadership will never be forgotten.