The Complete Analysis of a Polynomial Factorization Algorithm over Finite Fields

Philippe Flajolet, Xavier Gourdon, Daniel Panario

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

The complete analysis of a polynomial factorization algorithm over finite fields

Philippe Flajolet, Xavier Gourdon, Daniel Panario

No 3370
Mars 1998

THÈME 2

ISSN 0249-6399

The complete analysis of a polynomial factorization algorithm over finite fields

Philippe Flajolet, Xavier Gourdon, Daniel Panario

Thème 2 — Génie logiciel
et calcul symbolique
Projet Algo

Rapport de recherche— Mars 1998 — 28 pages

Abstract: A unified treatment of parameters relevant to factoring polynomials over finite fields is given. The framework is based on generating functions for describing parameters of interest and on singularity analysis for extracting asymptotic values. An outcome is a complete analysis of the standard polynomial factorization chain that is based on elimination of repeated factors, distinct degree factorization, and equal degree separation. Several basic statistics on polynomials over finite fields are obtained in the course of the analysis.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
Téléphone : (33) 01 39 63 55 11 – Télécopie : (33) 01 39 63 53 30

L'analyse complète d'un algorithme de factorisation de polynômes sur les corps finis

Résumé : Cet article propose un cadre unifié pour l'analyse des paramètres qui interviennent dans la factorisation de polynômes sur un corps fini. L'étude repose d'une part sur les séries génératrices utilisées pour décrire les principaux paramètres, d'autre part sur l'analyse de singularité qui fournit les informations asymptotiques correspondantes. Est ainsi obtenue une analyse complète d'une chaîne classique de factorisation fondée sur l'élimination des facteurs répétés, la factorisation en degré disticts, et la séparation des degrés égaux. Plusieurs statistiques de base relatives aux polynômes sur les corps finis résultent de cette analyse.

THE COMPLETE ANALYSIS OF A POLYNOMIAL FACTORIZATION ALGORITHM OVER FINITE FIELDS

PHILIPPE FLAJOLET, XAVIER GOURDON, AND DANIEL PANARIO

ABSTRACT. A unified treatment of parameters relevant to factoring polynomials over finite fields is given. The framework is based on generating functions for describing parameters of interest and on singularity analysis for extracting asymptotic values. An outcome is a complete analysis of the standard polynomial factorization chain that is based on elimination of repeated factors, distinct degree factorization, and equal degree separation. Several basic statistics on polynomials over finite fields are obtained in the course of the analysis.

1. INTRODUCTION

Factoring polynomials over finite fields intervenes in many areas like polynomial factorization over the integers [8, 31], cryptography [7, 33, 36], number theory [3], or coding theory [2]. The implications include finding complete partial fraction decompositions (a problem itself useful for symbolic integration), designing cyclic redundancy codes, computing the number of points on elliptic curves, and building arithmetic public key cryptosystems. In particular, the factorization of *random* polynomials over finite fields is needed in the index calculus method for computing discrete logarithms over finite fields [36].

This paper derives basic probabilistic properties of random polynomials over finite fields that are of interest in the study of polynomial factorization algorithms. We show that the main characteristics of random polynomial can be treated systematically by methods of "analytic combinatorics" based on the combined use of generating functions and of singularity analysis. Our object of study is the most basic factorization chain that is described in Fig. 1 and is close to standard implementations used in general purpose computer algebra systems [20]. In this paper, we provide a complete average-case analysis of this basic polynomial factorization chain which, despite its simplicity, does not appear to have been completely analysed previously.

The algorithm that we study need not be the fastest available at the moment; compare with [17, 18, 25, 43]. However, the discipline of analysing completely such an algorithm, which is in the line of Knuth's works [31], reveals parameters that are of general interest for polynomial factorization algorithms and for problems that deal with polynomials over finite fields at large. An illustration appears in [39] that deals with testing polynomial irreducibility along similar lines. On another register, Shoup [42] provides the average-case analysis of an algorithm he proposes. Shoup's highly interesting methods are based on estimates for the number of solutions of equations over finite fields and Weil's bounds; their scope is however quite different from the framework of this paper.

Algorithmic framework. The algorithmic problem that we address is as follows: given a monic univariate polynomial $f \in \mathbb{F}_q[x]$, find the complete factorization $f = f_1^{e_1} \cdots f_k^{e_k}$, where $f_1, \ldots, f_k$ are pairwise distinct monic irreducible polynomials and $e_1, \ldots, e_k$ are positive integers. The basic factorization chain (see Fig. 1) operates in three stages:

Date: February 26, 1998.

```
procedure factor(f : polynomial);
1.      a:=ERF(f);
2.      b:=DDF(a);
        F:=1;
3.      for k from 1 to n do
            F:=F.EDF(b[k],k);
        od;
4.      return(F.factor(f/a));
end;
```

FIGURE 1. The complete factorization chain.

ERF: *elimination of repeated factors* replaces a polynomial by a squarefree one that contains all the irreducible factors of the original polynomial with exponents reduced to 1;

DDF: *distinct-degree factorization* splits a squarefree polynomial into polynomials whose irreducible factors all have the same degree;

EDF: *equal-degree factorization* factors a polynomial whose irreducible factors have the same degree.

An often used variant of the first stage is:

SFF: *squarefree factorization* replaces a polynomial by a collection of squarefree ones that contain all the irreducible factors of the original polynomial with exponents reduced to 1.

As we shall see in Section 4 the difference in costs induced by the two versions, ERF and SFF, is marginal, while consideration of ERF greatly simplifies the whole analysis.

Computational model. We *fix* a finite field $\mathbb{F}_q$ with $q = p^m$ (p prime) and consider the polynomial ring $\mathbb{F}_q[x]$, see [20, 31, 34]. The probabilistic model assumes all q^n monic polynomials of degree n to be equally likely and all average-case analyses are expressed as asymptotic forms in n, the degree of the polynomial to be factored. The complexity model assumes that a basic field operation has cost $\mathcal{O}(1)$, the cost of a sum is $\mathcal{O}(n)$, and the cost of a product, a division or a gcd is $\mathcal{O}(n^2)$, when applied to polynomials of degree $\leq n$. For *dominant asymptotics*, we can freely restrict our attention to polynomial products and gcds. We take as $\tau_1 n^2$ the cost of multiplying two polynomials of degree less than n modulo a polynomial of degree n, and as $\tau_2 n^2$ the cost of a gcd between a polynomial of degree n and a polynomial of degree at most n. (There, τ_1 and τ_2 are system and implementation dependent constants.) What we have in mind is a general purpose factorization algorithm typically applied to polynomials of moderate size, for instance in a computer algebra system, where operations are implemented by quadratic algorithms. Similar studies could be conducted using FFT (fast Fourier transform) based algorithms.

An extended abstract of this paper has been presented at the ICALP'96 Conference [13].

2. Summary of results

A random polynomial of degree n is irreducible with probability tending to zero [2, 31], and has close to $\log n$ factors on average and with a high probability [5, 15]. Thus, the factorization of a random polynomial over a finite field is almost surely nontrivial. Other known results as well as our general framework based on generating functions and singularity analysis are presented in Section 3.

ERF. The first phase *ERF* of our factorization chain starts with the elimination of repeated factors, a simplified form of squarefree factorization described in Section 4. The ERF stage returns the *squarefree part* of the original polynomial, that is a polynomial in which each irreducible factor of the original polynomial appears exactly once. The remaining factors of the original polynomial form what we call the *non-squarefree part*. Theorem 1 quantifies this process and shows that up to smaller order terms, the expected cost is dominated by a single gcd of the polynomial f to be factored and its derivative f', so that it is $\mathcal{O}(n^2)$ on average. In a precise technical sense, most of the factorization cost results from the subsequent phases (DDF and EDF), since the non-squarefree part has average degree $\mathcal{O}(1)$.

DDF. The second phase *DDF* that is described in Section 5 splits the squarefree part a of the polynomial to be factored into a product $a = b_1 \cdot b_2 \cdots b_n$, where b_k is itself the product of the irreducible factors of a that have degree k. This phase is based on elementary properties of finite fields and is the one with the highest computational cost, namely $\mathcal{O}(n^3 \log q)$ on average. Theorem 5 provides a precise comparison of three strategies: the "naïve" rule, the "half-degree" rule and the "early abort" rule whose costs are found to be in the approximate proportion $1 :: \frac{3}{4} :: \frac{2}{3}$. Thus a savings of about one third results from controlling the DDF phase by the early abort strategy. These analyses involve the joint distribution of the largest and second largest irreducible factors of a random polynomial (Theorems 3 and 4). At the end of this phase, the factorization is complete with a probability ranging asymptotically between 0.56 and 0.67 (Theorem 6). In addition, the number of degree values such that more than one irreducible factor occurs is $\mathcal{O}(1)$, and the total degree passed to EDF is $\mathcal{O}(\log n)$ (Theorem 7).

EDF. The third phase *EDF* can be exactly analysed and its expected cost is comparatively small, being $\mathcal{O}(n^2 \log q)$ (see Lemma 4 and Theorem 9 for precise statements). For each nontrivial factor b_k, it involves a recursive refinement process again based on properties of finite fields. The analysis is close to that of digital trees known as "tries" [30] under a biased probability model.

Precise statements are given in the next sections with an explicit dependency on the field cardinality q, and some of them involve number-theoretic functions that can be both evaluated and estimated easily. A simplified picture is as follows. The ERF phase involves with high probability little more than a single polynomial gcd. The DDF phase of cost $\mathcal{O}(n^3 \log q)$ is the one that is most intensive computationally, where control by the "early-abort" strategy is expected to bring gains close to 36%. The last phase of EDF is executed less than 50% of the time and its cost is again small compared to that of DDF. A comparison between worst-case costs and average-case costs for each step is drawn in Section 8.

3. Basic methodology

This section gathers basic tools needed to analyse properties of random polynomials. It centers around the use of generating functions, either univariate or multivariate, whose functional relations reflect the algebraic decompositions of various classes of polynomials. The asymptotic analysis of coefficients of generating functions is then attained by means of singularities.

The results in this section are classical, but they are needed to set the stage for subsequent analyses. General references for this section are Chapter 3 of Berlekamp's book [2], the exercise section 4.6.2 of Knuth's book [31], and the paper by Flajolet and Odlyzko [14] for asymptotic methods.

3.1. **Generating functions.** We restrict our discussion to polynomials over a finite field $\mathbb{F}_q$. Let $\mathcal{I}$ be the collection of all monic irreducible polynomials. The formal identity

$$\frac{1}{1-\omega} = 1 + \omega + \omega^2 + \cdots$$

expresses arbitrary repetitions of the irreducible polynomial ω. The product, with repetitions allowed, of elements taken from $\mathcal{I}$ generates the collection $\mathcal{P}$ of all monic polynomials. In a similar way, the product of distinct elements taken from $\mathcal{I}$ generates the collection Q of all monic squarefree polynomials over $\mathbb{F}_q$. Thus, symbolically,

$$\mathcal{Q} = \prod_{\omega\in\mathcal{I}} (1+\omega), \qquad \text{and} \qquad \mathcal{P} = \prod_{\omega\in\mathcal{I}} (1-\omega)^{-1}. \tag{1}$$

In this context, $\mathcal{I}$ may itself be identified with

$$\mathcal{I} = \sum_{\omega\in\mathcal{I}} \omega.$$

Let z be a formal variable, and $|\omega|$ be the degree of $\omega \in \mathcal{I}$. The substitution $\omega \mapsto z^{|\omega|}$ in the above identity produces the generating function

$$I(z) = \sum_{\omega\in\mathcal{I}} z^{|\omega|} = \sum_n I_n z^n,$$

where I_n is the number of polynomials in $\mathcal{I}$ having degree n. Similarly, one has

$$\begin{aligned} Q(z) &= \prod_{\omega\in\mathcal{I}} (1+z^{|\omega|}) &&= \prod_{n=1}^{\infty} (1+z^n)^{I_n} \\ P(z) &= \prod_{\omega\in\mathcal{I}} (1-z^{|\omega|})^{-1} &&= \prod_{n=1}^{\infty} (1-z^n)^{-I_n}. \end{aligned} \tag{2}$$

The coefficients $Q_n = [z^n]Q(z)$ and $P_n = [z^n]P(z)$ evaluate to the number of monic squarefree polynomials of degree n, and to the number of monic polynomials of degree n, respectively. Obviously, $P_n = q^n$, and therefore we have

$$P(z) = \frac{1}{1-qz}. \tag{3}$$

The number of irreducible polynomials, I_n, is determined implicitly from the second relation of (2) that relates it to $P(z)$. A well-known process based on the Moebius inversion formula gives it in explicit form; see [6, p. 41], and Theorem 3.43 in [2, p. 84]. Indeed, taking logarithms and rearranging the sums leads to

$$\log \frac{1}{1-qz} = \sum_{k=1}^{\infty} \frac{I(z^k)}{k}, \quad \text{and} \quad \frac{q^n}{n} = \sum_{k|n} \frac{I_{n/k}}{k}. \tag{4}$$

The relation is then solved for I_n by means of Moebius inversion, which yields

$$I_n = \frac{1}{n} \sum_{k|n} \mu(k) q^{n/k}, \qquad \text{and} \qquad I(z) = \sum_{k=1}^{\infty} \frac{\mu(k)}{k} \log \frac{1}{1-qz^k}.$$

In particular, we have the important consequence

$$I_n = \frac{q^n}{n} + \mathcal{O}\left(\frac{q^{n/2}}{n}\right).$$

This identity shows that a fraction very close to $1/n$ of the polynomials of degree n is irreducible. As an aside, this result was first proven by Gauss for the case of prime fields. It appeared in his posthumous book [19]; see also [11], and [40].

Regarding the number of squarefree polynomials Q_n, each polynomial f factorizes uniquely as $f = s \cdot t^2$, where s is a squarefree polynomial and t is an arbitrary polynomial. (Separate the irreducible factors of f according to the parity of their exponents.) We thus have $P(z) = Q(z) \cdot P(z^2)$, so that

$$Q(z) = \frac{P(z)}{P(z^2)} = \frac{1 - qz^2}{1 - qz}.$$

Therefore, the number of squarefree polynomials of degree n in $\mathbb{F}_q[x]$ is

$$Q_n = \begin{cases} q^n & \text{if } n = 0, 1; \\ q^{n-1}(q-1) & \text{if } n \geq 2. \end{cases} \tag{5}$$

This result seems to have appeared for the first time in [6].

3.2. **Parameters.** We need extensions of the symbolic method in order to take care of characteristic parameters of polynomial factorization. Let Φ be a class of monic polynomials, and χ some integer–valued parameter on Φ. Then, the bivariate generating function

$$\phi(z, u) = \sum_{\omega \in \Phi} z^{|\omega|} u^{\chi(\omega)}$$

is such that the coefficient $[z^n u^k]\phi(z, u)$ represents the number of polynomials of degree n in Φ with χ–parameter equal to k. If χ is additive, meaning that $\chi(f \cdot g) = \chi(f) + \chi(g)$ for relatively prime $f, g \in \mathbb{F}_q[x]$, then the product decompositions of (2) generalize under the translation rule $\omega \mapsto z^{|\omega|} u^{\chi(\omega)}$. The technique of rearranging logarithms of infinite products employed in (4) is then useful in simplifying such expressions.

Bivariate generating functions contain all information related to the distribution of a parameter. Averages and standard deviations are obtained by successive differentiations of these bivariate generating functions with respect to u (the variable marking the parameter), and then by setting $u = 1$.

3.3. **Asymptotic analysis.** Generating functions encode exact informations on their coefficients. Furthermore, their behavior near their dominant positive singularity ("dominant" means in this context "of smallest modulus") is an important source of coefficient asymptotics.

The generating functions $f(z)$ to be studied in this paper are singular at $z = 1/q$ and most have there an isolated singularity. Consequently, their coefficients $f_n = [z^n]f(z)$ satisfy an estimate of the form $f_n \sim q^n \theta(n)$, where $\limsup |\theta(n)|^{1/n} = 1$ is a subexponential factor that reflects the nature of the singularity at $z = 1/q$ (see for instance [14], [37]). In common cases, an expansion near $z = 1/q$ of the form

$$f(z) = \frac{1}{(1 - qz)^\alpha} \left(\log \frac{1}{1 - qz} \right)^k (1 + o(1)) \tag{6}$$

translates into coefficients by the method known as singularity analysis [14]

$$f_n = [z^n]f(z) = q^n \frac{n^{\alpha-1}}{\Gamma(\alpha)} (\log n)^k (1 + o(1)). \tag{7}$$

The transition from (6) to (7) is ensured by transfer theorems that require analytic continuation of $f(z)$ outside its circle of convergence, a condition that is verified by inspection in most cases considered here. Another useful asymptotic coefficient extraction method is Darboux's

method [9, 24] whose principle is as follows: if an analytic function $f(z)$ defined in the closed disk $|z| \leq 1$ is k times continuously differentiable ($k \geq 0$) on $|z| = 1$, then its coefficients satisfy

$$[z^n]f(z) = o(1/n^k). \tag{8}$$

A recourse to Darboux's method is needed in Section 6, given the existence of natural boundaries for generating functions that occur there.

3.4. **The permutation model.** It is well-known that, as the cardinality q of the field goes to infinity (n staying fixed!), the joint distribution of the degrees of the irreducible factors in a random polynomial of degree n converges to the joint distribution of the lengths of cycles in a random permutation of size n. Generating functions for polynomials over a finite field $\mathbb{F}_q$ often have a dominant singularity at $z = 1/q$ while associated generating functions for permutations have a dominant singularity at $z = 1$. For parameters that only depend on the basic decomposition, this means that when q goes to infinity, generating functions of random polynomials at z/q converge to generating functions for permutations. For instance, the generating function of all monic polynomials, when normalized with the change of variable $z \mapsto z/q$, is

$$P\left(\frac{z}{q}\right) = \frac{1}{1-z} = \sum_{n=1}^{\infty} n! \frac{z^n}{n!},$$

which is the exponential generating function of all permutations. Similarly, we have

$$I\left(\frac{z}{q}\right) \underset{(q\to\infty)}{\to} \ln\frac{1}{1-z} = \sum_{n=1}^{\infty} (n-1)! \frac{z^n}{n!},$$

the exponential generating function of cyclic permutations. The relation between $P(z)$ and $I(z)$ that expresses the unique factorization property for polynomials reduces as $q \to \infty$ to

$$\frac{1}{1-z} = \exp\left(\ln\frac{1}{1-z}\right),$$

which is known [21] to express the unique decomposition of permutations into cycles.

This gives rise to a useful heuristic: probabilistic properties of polynomial factorization often have a shape resembling that of corresponding properties of the cycle decomposition of permutations to which they normally reduce as $q \to \infty$. An instance is mentioned in [24] in connection with the probability that a random polynomial admits factors of distinct degrees which, for large q and large n is found to approach $e^{-\gamma}$ (see Section 6 for more details).

4. Elimination of repeated factors (ERF)

The first step in the factorization chain of a polynomial is the elimination of its repeated factors. In characteristic zero, this is achieved by a gcd between f and its derivative f'. In $\mathbb{F}_q$, $q = p^m$ with p a prime number, additional control is needed in order to deal with pth powers whose derivatives are identically 0. The corresponding process for ERF is given in Fig. 2.

The first line of the algorithm collects in h one copy of each of the irreducible factors of f, except the ones whose multiplicity is a multiple of p. The while loop stores in g the factors whose multiplicity is a power of p, without eliminating their repetitions. The last part of the algorithm adds to h the factors in g with repetitions eliminated. The auxiliary computation of pth roots, $g^{1/p}$, is performed in the classical way [20, p. 344], using the identity $(a^p + b^p)^{1/p} = (a + b)$.

In order to obtain the complete factorization it is sufficient to recursively call ERF with the input polynomial f/h. There are other algorithms for the full squarefree factorization (see [31], Ex. 4.6.2.36); however, as Theorem 1 below shows, the reduction of degree induced by the additional computational effort is only $\mathcal{O}(1)$. Therefore, the whole analysis, as regards

```
procedure ERF(f : polynomial);
       g := gcd(f,f'); h := f/g; k := gcd(g,h);
       while k<>1 do  g := g/k; k := gcd(g,h)  od;
       if  g <> 1 then h := h*ERF(g^(1/p)) fi;
       return(h);
end;
```

FIGURE 2. The elimination of repeated factors algorithm (ERF).

dominant asymptotics, is not affected by the algorithm chosen for the first stage. We have thus chosen the elimination of repeated factors (ERF), a simpler algorithm than the full squarefree factorization (SFF).

Theorem 1. (i) *A random polynomial of degree $n \geq 2$ in $\mathbb{F}_q[x]$ has a probability $1 - 1/q$ to be squarefree.*

(ii) *The total degree ρ of the non-squarefree part of a random polynomial of degree n has an expected value that tends to*

$$C_q = \sum_{k \geq 1} \frac{k I_k}{q^{2k} - q^k},$$

and one has $C_q \sim 1/q$ as $q \to \infty$.

(*iii*) *The tail probabilities of ρ decay exponentially fast:*

$$\Pr(\rho(f) = k, \ |f| = n) = O\left(\left(\frac{2}{3}\right)^k\right).$$

PROOF. Part (i) is a classical result that we have already derived in (5). As for part (ii), the bivariate generating function of the degree of the non-squarefree part of monic polynomials in $\mathbb{F}_q[x]$ is, by the symbolic method of Section 3,

$$P(z,u) = \prod_{k \geq 1} \left(1 + \frac{z^k}{1 - u^k z^k}\right)^{I_k}.$$

The mean degree of the non-squarefree part is obtained by setting $u = 1$ in the derivative of $P(z,u)$ with respect to u and the asymptotic estimate follows by singularity analysis:

$$\begin{aligned} \left.\frac{\partial P(z,u)}{\partial u}\right|_{u=1} &= P(z) \cdot \sum_{k \geq 1} I_k \frac{z^{2k}}{1 - z^k} \\ &\underset{z \to 1/q}{\sim} \frac{1}{1 - qz} \sum_{k \geq 1} I_k \frac{q^{-2k}}{1 - q^{-k}}. \end{aligned}$$

The asymptotic value of C_q as $q \to \infty$ is obtained from there by the expansion $k\, I_k = q^k + \mathcal{O}(q^{k/2})$.

(*iii*) Consider the function $P(z, 3/2)$ and compare it with $Q(z)$, the generating function of square-free polynomials:

$$\frac{P(z,3/2)}{Q(z)} = \prod_{k \geq 1} \left(1 + \frac{z^{2k}(\frac{3}{2})^k}{(1 + z^k)(1 - (\frac{3}{2})^k z^k)}\right)^{I_k}.$$

The infinite product is convergent and analytic in a disk that properly contains $|z| \leq 1/q$. Thus, $P(z, 3/2)$ has a simple pole at $z = 1/q$, and by singularity analysis,

$$\frac{1}{q^n}[z^n]P(z, 3/2) \equiv \frac{1}{q^n} \sum_{|f|=n} \left(\frac{3}{2}\right)^{\rho(f)} = O(1),$$

which implies the exponential tail bound. ∎

Theorem 1 has important consequences for the recursive structure of the `factor` procedure. The exponential tail of Part (*iii*) implies that any polynomially bounded function of the nonsquare-free part stays of order $O(1)$ on average. In particular, the overall cost of the recursive calls (Step 4 in the top-level procedure) is $\mathcal{O}(1)$ on average. Alternative strategies that like SFF give the full squarefree factorization have asymptotically equivalent costs. Accordingly, the ERF phase has a cost dominated by that of its first gcd.

Theorem 2. *The expected cost of the ERF phase applied to a random polynomial of degree n is asymptotically that of a single gcd,*

$$\overline{\tau ERF}_n \sim \tau_2 n^2.$$

5. Distinct-degree factorization (DDF)

The second stage of the factorization chain is the distinct-degree factorization and it involves splitting a squarefree polynomial into polynomials whose irreducible factors all have the same degree. This means expressing the squarefree polynomial a in the form $b_1 \cdot b_2 \cdots b_n$ where b_k is the product of irreducible factors of degree k in a. The basic algorithm is described in Fig. 3 and it relies on the following fact (see [34], p. 91, Theorem 3.20).

Fact 1. *For $i \geq 1$, the polynomial $x^{q^i} - x \in \mathbb{F}_q[x]$ is the product of all monic irreducible polynomials in $\mathbb{F}_q[x]$ whose degree divides i.*

Three different stopping rules are considered:

— The *basic strategy* explores all the values $k = 1 \,..\, n$ and corresponds to the version given in the algorithm in Fig. 3.

— The *half-degree strategy* consists in stopping the DDF loop when $k = n/2$, since at that moment the remaining factor is either 1 or irreducible.

— The *early-abort strategy* stops the main loop of DDF as soon as $2k$ exceeds the degree of the remaining factor g. In this case, the remaining factor is by necessity irreducible.

The corresponding analyses are carried out in Subsection 5.2. They benefit from informations regarding the two largest irreducible factors of a random polynomial, a topic that we address first in Subsection 5.1.

5.1. Distribution of largest degrees of factors.

The distribution of the largest degree among the irreducible factors of a random polynomial over $\mathbb{F}_q$ underlies many problems dealing with polynomials over finite fields. For instance, information on this distribution is useful when computing discrete logarithms in order to discard polynomials that cannot be written in terms of smooth ones [36].

In the context of this paper, these properties provide insight on the stopping condition for the DDF stage. More specifically, we consider the two largest degrees $D_n^{[1]}$, $D_n^{[2]}$ of the distinct factors of a random polynomial of degree n in $\mathbb{F}_q$. The statements that follow are borrowed from [38]. The proofs are not given here as they are rather long and not in the spirit of the rest of the paper, while the results are only needed technically in the very last step of the analysis of the early abort rule (Lemma 3).

```
procedure DDF(a : polynomial);
[a is a monic squarefree polynomial]
       n := deg(a); g := a; h := x;
       for k:=1 to n do
1.         h:=h^q mod g;
2.         b[k]:=gcd(h-x,g);
3.         g:=g/b[k];        [a without factors of deg <=k]
4.         if b[k]<>1 then h:=h mod g fi;
       od;
       return(b[1].b[2]...b[n]);
end;
```

FIGURE 3. The "basic strategy" of the distinct-degree factorization algorithm (DDF).

Statistics on the largest degree of the irreducible factors of a random polynomial in $\mathbb{F}_q[x]$ have already been considered in the literature. Car [5] first obtained an asymptotic expression for the distribution function of $D_n^{[1]}$ in terms of the *Dickman function*. This function [10, 12, 44] is a classical number-theoretic function that describes the distribution of the largest prime divisor of a random integer, and it is defined as the unique continuous solution of the difference-differential equation

$$\rho(u) = 1 \qquad (0 \le u \le 1), \qquad\qquad u\rho'(u) = -\rho(u-1) \qquad (u > 1).$$

Gourdon [22, 23] has developed a new approach that also leads to local limit theorems. Theorems 3, 4 below, borrowed from [38], give the asymptotic distribution of the two largest degrees $D_n^{[1]}$, $D_n^{[2]}$ in the form of local limits, where the density functions are accessible via their Laplace transforms. A key rôle is played by the exponential integral function E defined by

$$E(a) = \int_a^{+\infty} \frac{e^{-v}}{v}\, dv.$$

Theorem 3. *The largest degree $D_n^{[1]}$ among the irreducible factors of a random polynomial of degree n over $\mathbb{F}_q$ satisfies*

$$\Pr(D_n^{[1]} = m) = \frac{1}{m} f\left(\frac{m}{n}\right) + \mathcal{O}\left(\frac{\log n}{m^2}\right),$$

where $f(\mu)$ is a continuous function that is defined by the inverse Laplace integral:

$$f(\mu) = \frac{1}{2\pi i} \int_{1-i\infty}^{1+i\infty} \frac{e^{-E(\mu h)}}{h}\, e^{(1-\mu)h}\, dh. \tag{9}$$

In other words, the largest degree is on average $\mathcal{O}(n)$, and the probability that its value equals m is about $1/m$ with a modulation factor that involves the Dickman function. (It can be verified by direct Laplace transform calculations that $f(\mu) = \rho(1/\mu - 1)$, with ρ the Dickman function.) The next theorem shows that similar estimates hold if a gap is imposed or if one considers the joint distribution of the largest two factors.

Theorem 4. *The two largest degrees $D_n^{[1]}$ and $D_n^{[2]}$ of the distinct factors of a random polynomial of degree n in $\mathbb{F}_q$ satisfy*

(i) *for* $0 \le m \le n$,

$$\Pr(D_n^{[1]} = m, D_n^{[2]} \le m/2) = \frac{1}{m} g_1\left(\frac{m}{n}\right) + \mathcal{O}\left(\frac{\log n}{m^2}\right),$$

where $g_1(\mu)$ *is defined by the inverse Laplace integral,*

$$g_1(\mu) = \frac{1}{2\pi i} \int_{1-i\infty}^{1+i\infty} \frac{e^{-E(\mu h/2)}}{h} e^{(1-\mu)h}\, dh;$$

(ii) *for* $0 \le m_2 < m_1 \le n$,

$$\Pr(D_n^{[1]} = m_1, D_n^{[2]} = m_2) = \frac{1}{m_1 m_2} g_2\left(\frac{m_1}{n}, \frac{m_2}{n}\right) + \mathcal{O}\left(\frac{\log n}{m_1 m_2^2}\right),$$

where $g_2(\mu_1, \mu_2)$ *is defined by*

$$g_2(\mu_1, \mu_2) = \frac{1}{2\pi i} \int_{1-i\infty}^{1+i\infty} \frac{e^{-E(\mu_2 h)}}{h} e^{(1-\mu_1-\mu_2)h}\, dh.$$

5.2. **Analysis of the distinct-degree factorization.** The main result of this section, Theorem 5, provides a quantitative comparison of the three stopping rules for the DDF algorithm. It is based itself on three lemmas corresponding to the three strategies.

We first fix some notation. The DDF algorithm in its basic version is specified in Fig. 3. The computation in step 1 is done by means of the classical *binary powering* method [31, p. 441-442] that leads to introducing two number-theoretic functions.

Definition 1. *The function* $\nu(q)$ *is the number of ones in the binary representation of* q. *The function* $\lambda(q)$ *is defined as*

$$\lambda(q) = \lfloor \log_2 q \rfloor + \nu(q) - 1, \tag{10}$$

and it represents the number of products needed to compute $h^q \bmod g$ *by binary powering.*

By the exponential tail result of Theorem 1, we need only consider the cost of DDF applied to the squarefree part a of the input polynomial f, and our subsequent analyses are all relative to the statistics induced by a random input f of degree n.

Let d_k denote the degree of polynomial g when the kth iteration of the main loop starts. The parameter d_k is also the sum of the degrees of the distinct factors of f with degree $\ge k$.

Theorem 5. *The expected cost of the DDF phase satisfies*

$$\overline{\tau DDF}_n^{(S)} \sim \mu^{(S)} (\lambda(q)\tau_1 + \tau_2)\, n^3 \qquad \lambda(q) = \lfloor \log_2 q \rfloor + \nu(q) - 1,$$

where $\mu^{(S)}$ *is a constant depending on the strategy* S:

$$\begin{aligned} \mu^{(B)} &= \frac{5}{12} \doteq 0.46667, \qquad \mu^{(HD)} = \frac{5}{16} = 0.3125, \\ \mu^{(EA)} &= \frac{5}{12} - \frac{1}{3}\int_0^\infty e^{-2x} \exp\left(-\int_x^\infty \frac{e^{-u}}{u}\, du\right) \frac{1-x^2}{x}\, dx \doteq 0.2668903307, \end{aligned}$$

for the basic (B), half-degree (HD) *or* early-abort (EA) *strategy.*

Proof. The cost of the basic DDF is $C_1 + C_2 + C_3 + C_4$, where C_j denotes the cost of line number j in the algorithm of Fig 3. Let $\overline{C_j}$ be the expectation of C_j. The mean number of irreducible factors of f is $\mathcal{O}(\log n)$, so that $\overline{C_3} + \overline{C_4} = \mathcal{O}(n^2 \log n)$; thus it suffices to consider $C_1 + C_2$.

The quantity d_k is the sum of the degrees of the distinct factors of f with degree $\geq k$. Then, the quantity $C_1 + C_2$ is equal to $(\lambda(q)\tau_1 + \tau_2)\sigma$, where

$$\sigma = \sum_{1 \leq k \leq N} d_k^2, \tag{11}$$

with N the stopping value for the DDF loop. We have $N = n$ for the basic strategy, $N = n/2$ for the half degree rule, $N = \max\{D_1/2, D_2\}$ for the early abort strategy, where D_1, D_2 are the largest and second largest degree of the distinct irreducible factors of the polynomial. The theorem follows from the next three lemmas. ∎

The analysis of the basic strategy only involves an additive parameter of polynomial factorizations. It is thus dealt with directly by bivariate generating functions and singularity analysis. The estimate also serves (by difference) in the analysis of the other two strategies.

Lemma 1. *The expected value of σ in the basic strategy satisfies, as $n \to \infty$,*

$$\overline{\sigma}_n^{(B)} \sim \frac{5}{12}\, n^3.$$

Proof. The parameter d_k is by definition the sum of the degrees of the distinct factors of f with individual degree $\geq k$ and the paremeter $\sigma^{(B)}$ of (11) is $\sigma^{(B)} = \sum_{k=1}^{n} d_k^2$. The bivariate generating function of d_k is, by the basic decompositions,

$$P_k(z,u) = \prod_{j<k} \left(\frac{1}{1-z^j}\right)^{I_j} \prod_{j \geq k} \left(1 + u^j \frac{z^j}{1-z^j}\right)^{I_j}.$$

The expected value of $\sigma^{(B)}$ is then given by

$$\overline{\sigma}_n^{(B)} = \frac{1}{q^n}[z^n]\xi(z), \qquad \xi(z) = \sum_{k \geq 1} \left(\frac{\partial^2 P_k}{\partial u^2}(z,u) + \frac{\partial P_k}{\partial u}(z,u)\right)\Big|_{u=1}.$$

The basic relation (2) and a computation of $\xi(z)$ by logarithmic derivatives imply that $\xi(z) = P(z)\,(\xi_1(z) + \xi_2(z))$ where

$$\xi_1(z) = \sum_{j \geq 1} j^3 I_j (z^j - z^{2j}) \qquad \text{and} \qquad \xi_2(z) = \sum_{k \geq 1} \left(\sum_{j \geq k} j I_j\, z^j\right)^2.$$

The function $\xi_1(z)$ is a simple variant of $I(z)$, namely

$$\xi_1(z) = \left(z\frac{d}{dz}\right)^3 \left(I(z) - \frac{1}{8} I(z^2)\right).$$

Thus, $\xi_1(z)$ has its dominant singularity at $z = 1/q$ and near this point,

$$P(z)\xi_1(z) \sim \frac{2}{(1-qz)^4}.$$

Singularity analysis then yields $[z^n]P(z)\xi_1(z) \sim q^n\, n^3/3$.

We next turn to $\xi_2(z)$. The estimate $jI_j = q^j + \mathcal{O}(q^{j/2})$ entails, for $|z| < 1 + \eta$,

$$\sum_{j \geq k} j I_j \left(\frac{z}{q}\right)^j = \frac{z^k}{1-z} + S_k(z), \quad S_k(z) = \mathcal{O}\left(\frac{z^k}{q^{k/2}}\right),$$

so that

$$\xi_2\left(\frac{z}{q}\right) = \frac{z^2}{1+z}\,\frac{1}{(1-z)^3} + S(z), \quad S(z) = \sum_{k \geq 1} \left(\frac{2z^k S_k(z)}{1-z} + S_k(z)^2\right).$$

The bounds satisfied by $S_k(z)$ make $S(z)$ regular beyond the unit disk. Thus, singularity analysis can be applied to the function $P(z/q)\xi_2(z/q)$. There are two dominant singularities at 1 and -1, but the one at $z=1$ has greater weight. The resulting estimate, as $z \to 1$,

$$P\left(\frac{z}{q}\right)\xi_2\left(\frac{z}{q}\right) \sim \frac{1}{2(1-z)^4},$$

implies that the nth coefficient is asymptotic to $n^3/12$. Thus, finally, $\overline{\sigma}_n^{(B)} \sim (1/3+1/12)n^3 = 5n^3/12$. ∎

The analysis of the half-degree rule is related to Theorem 3 since it involves the distribution of $D^{[1]}$. However, it only depends on areas where the Dickman function trivializes so that a direct argument applies.

Lemma 2. *The expected value of σ for the half-degree rule satisfies, as $n \to \infty$,*

$$\overline{\sigma}_n^{(HD)} \sim \frac{5}{16}\, n^3.$$

Proof. We now have $\sigma^{(HD)} = \sum_{k \le n/2} d_k^2$. It suffices to consider the difference $\sigma' = \sigma^{(B)} - \sigma^{(HD)} = \sum_{n/2<k\le n} d_k^2$. If the largest degree $D_n^{[1]}$ satisfies $D_n^{[1]} \le n/2$, we have $\sigma' = 0$. Otherwise we have $\sigma' = (D_n^{[1]} - \lfloor n/2 \rfloor)\,(D_n^{[1]})^2$ since there can be only one factor of degree larger than $n/2$, namely $D_n^{[1]}$. Thus, the mean value of σ' is given by

$$\overline{\sigma'}_n = \sum_{n/2<k\le n} \Pr(D_n^{[1]} = k)\left(k - \left\lfloor \frac{n}{2} \right\rfloor\right) k^2. \tag{12}$$

By the symbolic method of Section 3, the generating function of polynomials for which all factors have degree $\le m$ is

$$\chi_m(z) = \prod_{k=1}^{m} \left(\frac{1}{1-z^k}\right)^{I_k}. \tag{13}$$

Thus, the generating function of polynomials for which $D_n^{[1]} = m$ is

$$\phi_m(z) = \chi_m(z) - \chi_{m-1}(z) = \chi_m(z)\left(1 - (1-z^m)^{I_m}\right), \tag{14}$$

and the probability $\Pr(D_n^{[1]} = k)$ is related to this generating function by

$$\Pr(D_n^{[1]} = k) = \frac{1}{q^n}[z^n]\phi_k(z).$$

When $k > n/2$, the nth coefficient of $\phi_k(z)$ is obtained from

$$\phi_k(z) = P(z)\left(1-(1-z^k)^{I_k}\right)\prod_{j>k}(1-z^j)^{I_j} = P(z)\left(I_k\, z^k + \mathcal{O}(z^{n+1})\right)$$

which entails $\Pr(D_n^{[1]} = k) = I_k/q^k \sim 1/k$ for $n/2 < k \le n$. Plugging this estimate into (12) gives $\overline{\sigma'}_n \sim \frac{5}{48}\, n^3$. Thus $\overline{\sigma}_n^{(HD)} = \overline{\sigma}_n^{(B)} - \overline{\sigma'}_n \sim \frac{5}{16}\, n^3$. ∎

The early-abort strategy needs to be handled in a more technical way. There is a striking parallel with the analysis of integer factoring given by Knuth and Trabb-Pardo [32].

Lemma 3. *The expected value of σ in the early-abort rule satisfies, as $n \to \infty$,*

$$\overline{\sigma}_n^{(EA)} \sim \delta\, n^3,$$

where

$$\delta = \frac{5}{12} - \frac{1}{3}\int_0^\infty e^{-2x} \exp\left(-\int_x^\infty \frac{e^{-u}}{u}\,du\right)\frac{1-x^2}{x}\,dx \doteq 0.2668903307. \tag{15}$$

PROOF. (*i*) *Algebra.* As above, we denote by D_1 the degree of the largest irreducible factor of f, and by D_2 the degree of the second largest irreducible factor of f (set $D_2 = 0$ if f is irreducible). The iteration is now aborted at step $k_0 = \max\{\lfloor D_1/2\rfloor, D_2\}$ and $\sigma^{(EA)} = \sum_{k\le k_0} d_k^2$. Consider the difference

$$\sigma'' = \sigma^{(B)} - \sigma^{(EA)} = \sum_{\max\{\lfloor D_1/2\rfloor, D_2\} < k \le n} d_k^2.$$

We need to prove the mean-value estimate $\overline{\sigma''}_n \sim (\frac{5}{12} - \delta)n^3$.

We have $\sigma'' = (D_1 - \lfloor D_1/2\rfloor)D_1^2$ if $D_1/2 \ge D_2$, and $\sigma'' = (D_1 - D_2)D_1^2$ if $D_1/2 < D_2$. Thus, the mean value of σ'' is

$$\begin{aligned}\overline{\sigma''}_n &= \sum_{D_1=1}^{n} D_1^2\left(D_1 - \left\lfloor\frac{D_1}{2}\right\rfloor\right)\Pr(D_n^{[1]} = D_1, D_n^{[2]} \le D_1/2)\\ &+ \sum_{D_2 < D_1 \le 2D_2} D_1^2(D_1 - D_2)\Pr(D_n^{[1]} = D_1, D_n^{[2]} = D_2).\end{aligned} \tag{16}$$

Hence, the generating function $\Psi(z)$ of the cumulated values of the parameter σ'' is expressible in terms of the generating function $\widetilde{\phi}_{D_1}(z)$ of polynomials for which $D_n^{[1]} = D_1, D_n^{[2]} \le D_1/2$, and the generating function $\phi_{D_1,D_2}(z)$ of polynomials for which $D_n^{[1]} = D_1, D_n^{[2]} = D_2$. By symbolic methods, the generating function of polynomials for which $D_n^{[1]} = m$ and $D_n^{[2]} \le m/2$ is

$$\widetilde{\phi}_m(z) = \chi_{\lfloor m/2\rfloor}(z)\,\frac{I_m z^m}{1-z^m}, \tag{17}$$

with $\chi_m(z)$ defined in (13). In the same way, the generating function of polynomials with $D_1^{[n]} = m_1$ and $D_2^{[n]} = m_2$, $m_2 < m_1$ is

$$\phi_{m_1,m_2}(z) = \phi_{m_2}(z)\,\frac{I_{m_1} z^{m_1}}{1-z^{m_1}}, \tag{18}$$

with $\phi_m(z)$ defined in (14). Thus,

$$\Psi(z) = \sum_{D_1}\left(D_1 - \left\lfloor\frac{D_1}{2}\right\rfloor\right)D_1^2\,\widetilde{\phi}_{D_1}(z) + \sum_{D_2 < D_1 \le 2D_2}(D_1 - D_2)D_1^2\,\phi_{D_1,D_2}(z).$$

(*ii*) *Analysis.* The analysis of the generating function $\Psi(z)$ near the positive singularity $z = 1/q$ is done by approximating sums with integrals (Euler-Maclaurin summation). The following fact summarizes what is needed.

Fact 2. *The remainders of the logarithm* $r_m(z) = \sum_{k>m} z^k/k$ *are approximable in terms of the exponential integral,*

$$r_m(e^{-h}) = E(mh) + \mathcal{O}\left(\frac{1}{m}\right), \tag{19}$$

where the big-Oh error term is uniform with respect to $h > 0$.

Justifying this for any $h > 0$ only requires the Euler Maclaurin formula and "subtraction of singularities",

$$\begin{aligned} r_m(e^{-h}) &= \int_h^{+\infty} \left(\sum_{k>m} e^{-ku} \right) du = \int_{mh}^{+\infty} e^{-v} \frac{v/m}{e^{v/m}-1} \frac{dv}{v} \\ &= E(mh) + R_m(mh), \qquad R_m(u) = \frac{1}{m} \int_u^{+\infty} e^{-v} \eta\left(\frac{v}{m}\right) dv, \end{aligned}$$

with $\eta(z) = 1/(e^z - 1) - 1/z$ that is continuous over the positive half-line.

The estimate of Equation (19) applied to $\chi_m(z)$ and $\phi_m(z)$ entails

$$\chi_m\left(\frac{e^{-h}}{q}\right) = \frac{e^{-E(mh)+\mathcal{O}(1/m)}}{1-e^{-h}}, \qquad \phi_m\left(\frac{e^{-h}}{q}\right) = \frac{e^{-E(mh)+\mathcal{O}(1/m)}}{1-e^{-h}} \frac{e^{-mh}}{m}. \tag{20}$$

When $z = e^{-t}/q$ with $t > 0$, the evaluation (20) together with the expressions (17) and (18) of the intervening generating functions give

$$\widetilde{\phi}_{D_1}(z) \approx \frac{e^{-tD_1}}{D_1} \frac{e^{-E(\lfloor D_1/2 \rfloor t)}}{t}, \qquad \phi_{D_1,D_2}(z) \approx \frac{e^{-(D_1+D_2)t}}{D_1 D_2} \frac{e^{-E(D_2 t)}}{t}.$$

Approximating sums by integrals, when $z = e^{-t}/q$ with $t \to 0^+$, yields

$$\Psi(z) \sim \frac{1}{2t} \int_0^{\infty} x^2 e^{-tx} e^{-E(xt/2)}\, dx + \frac{1}{t} \int_{y<x<2y} (x-y) x^2 \frac{e^{-(x+y)t}}{xy} e^{-E(ty)}\, dx\, dy.$$

The change of variables $tx \mapsto x$ and $ty \mapsto y$, rephrases the double integral as

$$\Psi\left(\frac{e^{-t}}{q}\right) \sim \frac{4}{t^4} \int_0^{\infty} x^2 e^{-2x} e^{-E(x)}\, dx + \frac{1}{t^4} \int_0^{\infty} e^{-2y} \frac{2 + y - e^{-y}(2 + 3y + 2y^2)}{y}\, dy.$$

These integrals simplify under partial integration, and one finds

$$\Psi\left(\frac{e^{-t}}{q}\right) \sim \frac{c}{t^4} \qquad (t \to 0^+), \qquad c = \frac{5}{2} - 6\delta. \tag{21}$$

(*iii*) *Coefficients.* The asymptotic form (21) of $\Psi(z)$ as $t \to 0$ is consistent with the assertion that

$$[z^n]\Psi(z) \sim \frac{c}{6} q^n n^3. \tag{22}$$

However, an asymptotic estimate like (21) is confined to the vicinity of the real line, since it can be proved that a function like $\Psi(z)$ admits its circle of convergence as a natural boundary. Thus singularity analysis cannot be applied. (A Tauberian theorem could be tried, but Tauberian side conditions appear to be delicate to establish.) We then proceed instead by an Abelian argument that is based on a direct proof of existence of the limit

$$\varpi := \lim_{n \to \infty} \frac{1}{q^n n^3} [z^n]\Psi(z). \tag{23}$$

Indeed, Formula (16) together with Theorem 4, guarantees the existence of the limit in (23) since

$$\begin{aligned} \overline{\sigma''}_n &\sim \sum_{D_1=1}^{n} D_1 \left(D_1 - \left\lfloor \frac{D_1}{2} \right\rfloor \right) g_1\left(\frac{D_1}{n}\right) + \sum_{D_2 < D_1 \le 2D_2} D_1 \left(\frac{D_1}{D_2} - 1 \right) g_2\left(\frac{D_1}{n}, \frac{D_2}{n}\right) \\ &\sim \left[\int_0^1 \frac{x^2}{2} g_1(x)\, dx + \int_0^1 x \left(\int_{x/2}^{x} \left(\frac{x}{y} - 1 \right) g_2(x,y) dy \right) dx \right] n^3, \end{aligned}$$

where, in the second line, Riemann sums have been approximated by integrals. This is sufficient to conclude on the existence of ϖ in (23). This value can then be identified with $c/6$ in (21) and (22). ■

The constant δ in the above proof is a close relative of the famous Golomb constant that intervenes in the expectation of the longest cycle in a random permutation [41].

The global savings of the early abort strategy is thus of 36% compared to the basic strategy, and of 15% compared to the half-degree rule. The expected cost of DDF is $\mathcal{O}(n^3 \log q)$ and this cost dominates in the whole factorization chain.

6. The output configuration of DDF

The DDF procedure does not completely factor a polynomial that has different irreducible factors of the same degree. However, as shown by the following results, "most" of the factoring has been completed after DDF. First, the DDF procedure is a complete factorization with asymptotic probability greater than 1/2 (Theorem 6). Next, the number of calls to the subsequent phase of EDF, that is to say the number of degree values for which more than one factor occurs, is only $\mathcal{O}(1)$, and the sum of the degrees where this happens (the total degree of the fragments passed to EDF) is $\mathcal{O}(\log n)$. However, this total degree has a fairly large variability so that the cost of EDF (to be analysed in the next section) is comparatively small but not completely negligible. Theorem 7 quantifies some of these phenomena. They are established here by means of a hybridization of singularity analysis and Darboux's method, a general technique that we explain in some detail when we first encounter it in the next theorem.

Theorem 6. *The asymptotic probability for the distinct-degree factorization to be the complete factorization is*

$$c_q = \prod_{k\geq 1} \left(1 + \frac{I_k}{q^k - 1}\right) (1 - q^{-k})^{I_k}.$$

$c_2 \doteq 0.6656$, $c_3 \doteq 0.6123$, $c_5 \doteq 0.5861$, $c_{47} \doteq 0.5635$, $c_{257} \doteq 0.5618$, $c_\infty = e^{-\gamma} \doteq 0.5614$, *where* γ *is Euler's constant.*

Proof. We start with the analysis of the permutation model, as this illustrates a "bare-bones" version of the method. By remarks above, this corresponds to the limit case $q = \infty$. The probability that a permutation of length n has all its cycles of different lengths is $[z^n]F(z)$, where the generating function $F(z)$ is susceptible to a variety of expressions obtained by the technique of convergence factors:

$$\begin{aligned} F(z) &:= \prod_{k=1}^{\infty} \left(1 + \frac{z^k}{k}\right) \\ &= e^{-z} \frac{1+z}{1-z} \prod_{k=2}^{\infty} \left(1 + \frac{z^k}{k}\right) e^{-z^k/k} \\ &= \left(\frac{1+z}{1-z} \exp\left(-\frac{1}{2}\mathrm{Li}_2\left(z^2\right)\right)\right) \cdot \left(e^{-z+z^2/2} \prod_{k=2}^{\infty} \left(1 + \frac{z^k}{k}\right) e^{-z^k/k + z^{2k}/(2k^2)}\right) \\ &= S(z) \cdot R(z). \end{aligned} \tag{24}$$

Here $\mathrm{Li}_2(z) := \sum_{k\geq 1} z^k/k^2$ is the classical *dilogarithm* function. The first factor, $S(z)$, in the bottom equality of (24) satisfies the conditions of singularity analysis, while the second one, $R(z)$ is continuously differentiable (of class $\mathcal{C}^1$) on the closed unit disc $\overline{D}$, since it is of the form $e^{r(z)}$ where the coefficient $[z^n]r(z)$ is $\mathcal{O}(n^{-3})$.

We thus have a situation where the generating function of interest is the product of a singular part $S(z)$ that satisfies strong analyticity properties outside of $z = \pm 1$, and of a function $R(z)$ of the Darboux type that is smooth on the closed unit disc $\overline{D}$. We only need to justify the fact that dominant asymptotics of the coefficients $[z^n]F(z)$ can be extracted "as though" $R(z)$ was analytic on $\overline{D}$.

The local expansions of $S(z)$ at its singularities ± 1 are readily found to be

$$\begin{aligned} S_{+1}(z) &= 2e^{-\zeta(2)/2}\left(\frac{1}{1-z} - \log 2 + \frac{1}{2} - \log(1-z) + \mathcal{O}((1-z)\log^2(1-z))\right) \\ S_{-1}(z) &= \frac{1}{4}e^{-\zeta(2)/2}\,(2(z+1) + \mathcal{O}((z+1)\log(z+1)), \end{aligned}$$

with the error terms being $\mathcal{C}^0$ on $\overline{D}$. In summary, we have found that

$$F(z) = \left(2e^{-\zeta(2)/2}\frac{1}{1-z} + \log(1-z) + t(z)\right)\cdot R(z), \tag{25}$$

for some $t(z)$ that is $\mathcal{C}^0$, and $R(z)$ that is $\mathcal{C}^1$ on $\overline{D}$.

The expansion $R(z) = R(1) + U(z)(z-1)$ with a function $U(z)$ that is $\mathcal{C}^0$, can then be inserted in (25). Darboux's method applies to the resulting form for $F(z)$ (see the discussion relative to Equation (8) in Section 3.3), and one has

$$[z^n]F(z) = 2R(1)e^{-\zeta(2)/2} + o(1), \qquad R(1) := e^{-1/2}\prod_{k=2}^{\infty}\left(1+\frac{1}{k}\right)e^{-1/k+1/(2k^2)}.$$

This finally simplifies using the infinite product formula for $\Gamma(1)$:

$$[z^n]F(z) = \prod_{k=1}^{\infty}\left(1+\frac{1}{k}\right)e^{-1/k} + o(1) = e^{-\gamma} + o(1).$$

This result has been already established by Greene and Knuth [24] by means of a Tauberian argument combined with bootstrapping. The method used here is in contrast a hybrid of singularity analysis and of Darboux's method. It can be employed to derive complete asymptotic expansions, with roots of unity that intervene in successive asymptotic terms corresponding to smaller and smaller singularity weights. (This leads to fluctuating terms involving successive roots of unity.)

We next turn to the case of a finite field of fixed cardinality q to which the same principles apply. The generating function of polynomials with irreducible factors all of distinct degrees is, by standard decomposition formulæ,

$$F(z) = \prod_{k\geq 1}\left(1 + I_k(z^k + z^{2k} + z^{3k} + \cdots)\right) = \prod_{k\geq 1}\left(1 + I_k\frac{z^k}{1-z^k}\right). \tag{26}$$

An equivalent form that reveals the pole-like singularity at $z = 1/q$ is obtained by multiplying each term of the product (26) by $(1-z^k)^{I_k}$,

$$F(z) = \frac{1}{1-qz}\prod_{k\geq 1}\left(1 + I_k\frac{z^k}{1-z^k}\right)(1-z^k)^{I_k}.$$

Thus, as $z \to q^{-1}$, we have

$$F(z) \sim \frac{c_q}{1-qz}, \qquad c_q = \prod_{k=1}^{\infty}\left(1 + I_k\frac{1}{q^k-1}\right)(1-q^{-k})^{I_k}, \tag{27}$$

and the preceding discussion applies, with the rôle of the dilogarithm function now played by

$$\Lambda_2(z) = \sum_{k=1}^{\infty} \left(\frac{I_k}{q^k - 1}\right)^2 z^k.$$

The hybrid method then yields,

$$[z^n] F\left(\frac{z}{q}\right) = c_q + o(1),$$

which, by (27), is our statement. ∎

Theorem 6 was obtained by Knopfmacher and Warlimont [28] and independently by the authors in [13]. The methods of [13] and the present paper are however rather different from those of [28]. The paper [28] uses elementary techniques and derives constructive bounds. The methods developed here are less constructive but they are geared towards full asymptotic expansions and have been successfully used by Gourdon [22] to solve the Golomb-Knuth conjecture [29, Ex. 1.3.3.23] regarding the expectation of maximal cycle lengths in random permutations.

Theorem 7. *The number N_0 of degree values for which there is more than one irreducible factor produced by DDF has an average that is asymptotic to the constant*

$$\mu_0 = \sum_{k\geq 1} (1 - q^{-k})^{I_k} \left((1 - q^{-k})^{-I_k} - 1 - \frac{I_k q^{-k}}{1 - q^{-k}} \right).$$

The total degree N_1 of the corresponding polynomials has expectation $\log n + \mathcal{O}(1)$. and standard deviation of order $\sqrt{n}$.

Proof. Given a family $\mathcal{F}$ of elements, the expression

$$\prod_{\omega \in \mathcal{F}} \frac{1}{1 - \omega}$$

formally generates all multisets. The expression

$$1 + \sum_{\omega \in \mathcal{F}} \frac{\omega}{1 - \omega} + u \left(\prod_{\omega \in \mathcal{F}} \frac{1}{1 - \omega} - 1 - \sum_{\omega \in \mathcal{F}} \frac{\omega}{1 - \omega} \right)$$

formally generates all multisets each affected by a coefficient of u if there are different elements, and 1 otherwise. This applies to the class of irreducible polynomials of each degree n, taking $\mathcal{F} = \mathcal{I}_n$. Thus, the bivariate generating function of the number N_0 of degree values for which there are repeated elements is

$$P_0(z, u) = \prod_{k\geq 1} \left(1 + \frac{I_k z^k}{1 - z^k} + u \left((1 - z^k)^{-I_k} - 1 - \frac{I_k z^k}{1 - z^k} \right) \right). \tag{28}$$

The logarithmic derivative with respect to u satisfies

$$\frac{P_0'(z, 1)}{P_0(z, 1)} = \sum_{k\geq 1} (1 - z^k)^{I_k} \left((1 - z^k)^{-I_k} - 1 - \frac{I_k z^k}{1 - z^k} \right). \tag{29}$$

By our general discussion of the hybrid singularity analysis and Darboux method, the quantity N_0 has an expectation that is asymptotic to the limit μ_0 of $P_0'(z, 1)/\ P_0(z, 1)$ as $z \to 1/q$. This quantity is thus nothing but the value of the right hand side of (29) at $z = 1/q$.

For the sum N_1 of the degrees of these polynomials, an adaptation of (28) yields the bivariate generating function,

$$P_1(z,u) = \prod_{k\geq 1}\left(\left(1+u^k\frac{z^k}{1-z^k}\right)^{I_k} - (u^k-1)I_k\,\frac{z^k}{1-z^k}\right).$$

We only discuss briefly the first moment of N_1. The mean value is $q^{-n}[z^n]R(z)$, where $R(z)$ equals $P_1'(z,u)|_{u=1}$. Thanks to the expansion $kI_k = q^k + \mathcal{O}(q^{k/2})$, near $z = 1/q$, $R(z)$ is asymptotic to $(1-qz)^{-1}\log(1-qz)^{-1}$. Thus, the expectation of N_1 taken over polynomials of degree n is $q^{-n}[z^n]R(z) \sim \log n$. The second factorial moment of N_1 is obtained by a further differentiation of $P_1(z,u)$. ■

The analysis of N_1 in Theorem 7 was given in [13] and Knopfmacher [26] has independently obtained an estimate of the first two moments of N_0.

It should be clear that the hybrid asymptotic method has great flexibility. As a final illustration, we discuss a question of von zur Gathen and consider the quantity $\widetilde{N}$ that is the largest degree for which two or more factors occur. The generating function of polynomials such that $\widetilde{N} \leq r$ is in this case

$$F^{(r)}(z) = \prod_{k\leq r}(1-z^k)^{-I_k}\prod_{k>r}\left(1+I_k\frac{z^k}{1-z^k}\right).$$

Thus, the probability of $\widetilde{N} \leq r$ is, for large degree n and fixed r, asymptotic to

$$c_q^{(r)} = \prod_{k>r}(1-q^{-k})^{-I_k}\left(1+I_k\frac{q^{-k}}{1-q^{-k}}\right), \tag{30}$$

and for large field cardinalities, these constants tend to

$$c_\infty^{(r)} = e^{-\gamma}\prod_{k\leq r}\left(1+\frac{1}{k}\right)^{-1}e^{1/k}. \tag{31}$$

We have $1-c_q^{(r)} = \mathcal{O}(1/r)$ for all fixed q, some representative values with $q=\infty$ being:

$$c_\infty \doteq 0.5614,\ c_\infty^{(1)} \doteq 0.7631,\ c_\infty^{(2)} \doteq 0.8387,\ c_\infty^{(5)} \doteq 0.9179,\ c_\infty^{(10)} \doteq 0.9549.$$

Thus, (30) and (31) give the following simplified picture in the asymptotic limit (n and q large). A random polynomial has a small number, $\mathcal{O}(1)$, of "colliding" degrees; the largest colliding degree has a (defective) probability distribution with a tail that decays like $\mathcal{O}(1/r^2)$. Because of this slow tail decay, the largest colliding degree alone has a first moment that is $\mathcal{O}(\sum_r r^{-1}) = \mathcal{O}(\log n)$, and a second moment that is $\mathcal{O}(\sum_r 1) = \mathcal{O}(n)$. These observations are seem to be consistent with the facts asserted in Theorem 7.

7. Equal-degree factorization (EDF)

After the first two stages of the general algorithm, the factorization problem has been eventually reduced to factoring a collection of monic squarefree polynomials b_k all of whose irreducible factors have the same (known) degree k. The third step in the factorization process, the equal-degree factorization algorithm (EDF), focuses on polynomials with this special form. Our reference chain uses the classical Cantor-Zassenhaus algorithm [4] for this purpose. The analysis combines a recursive partitioning problem akin to digital trees —also known as "tries" [30]— together with estimates on the degrees of irreducible factors of random polynomials [27]. The net result is that the global cost of EDF is quadratic, a sharp contrast with the cubic cost of

```
procedure EDF(b : polynomial, k : integer);
[b is a product of irreducibles of degree k]
       if degree(b) <= k then return(b) fi;
       h := randpoly(degree(b)-1);
1.     a := h^((q^k-1)/2)-1 mod b;
2.     d := gcd(a,b);
       return(EDF(d,k).EDF(b/d,k));
end;
```

FIGURE 4. The equal-degree factorization algorithm (EDF).

DDF. For convenience, we first assume that q is odd, and relegate to Section 7.4 the case of a characteristic equal to 2.

The EDF algorithm is described in Fig. 4, and we briefly recall its principle here. Let b be a polynomial that is a product of j irreducible factors $f_1, \ldots, f_j$, each of degree k. The Chinese remainder theorem implies the ring homomorphism,

$$\mathbb{F}_q[x]/(b) \cong \mathbb{F}_q[x]/(f_1) \times \cdots \times \mathbb{F}_q[x]/(f_j),$$

and a random element h of $\mathbb{F}_q[x]/(b)$ is associated to a j-tuple $(h_1, \ldots, h_j)$, where each h_i is a random element of $\mathbb{F}_q[x]/(f_i)$.

Now comes the splitting principle that eventually makes it possible to isolate the various f_i. Since each f_i is irreducible, the multiplicative group of each component $\mathbb{F}_q[x]/(f_i)$ is a field isomorphic to $\mathbb{F}_{q^k}$. Such a group being cyclic, there are the same number $(q^k-1)/2$ of squares and nonsquares. The test $h_i^{(q^k-1)/2} = 1$ discriminates the squares in this multiplicative group. Thus, taking a random h and computing $a := h^{(q^k-1)/2} - 1 \bmod b$, we have that $\gcd(a, b)$ "extracts" the product of all the f_i for which h is a square in $\mathbb{F}_q[x]/(f_i)$.

From the probabilistic point of view, a component h_i that is random in $\mathbb{F}_q[x]/(f_i)$ has probability $\alpha = \frac{1}{2} - \frac{1}{2q}$ of being discriminated by the gcd test and the dual probability, $\beta = \frac{1}{2} + \frac{1}{2q}$ of being a nonsquare. The (small) difference between α and β is accounted for by the possibility of having noninvertible components.

In summary, irreducible factors of each degree can be extracted successively. For each degree where two or more such factors occur, the recursive splitting process will be launched. Then, the analysis of the complete EDF phase (Section 7.3) requires a purely combinatorial analysis of what takes place at each degree (Section 7.2) combined with an estimate of the probability that there are j irreducible factors of degree k in a random polynomial of degree n. These probabilities give interesting information on random polynomials and have been obtained by Knopfmacher and Knopfmacher [27] whose results we recall in Section 7.1.

7.1. **Irreducible factors of each degree.** Let $\kappa_n(k)$ be the random variable counting the number of distinct irreducible factors of degree k in a random polynomial of degree n. We consider here k as fixed. The corresponding probability distribution can be computed by the decomposition techniques of Section 2, see [27]. The bivariate generating function for the number of irreducibles of degree k is, by the basic decomposition,

$$\begin{aligned} Q_k(z,u) &= \left(1 + u\frac{z^k}{1-z^k}\right)^{I_k} \prod_{\ell \neq k} (1-z^\ell)^{-I_\ell} \\ &= \frac{1}{1-qz}(1+(u-1)z^k)^{I_k}. \end{aligned}$$

The asymptotics as $n \to \infty$ are derived from the polar singularity at $z = 1/q$: the probability generating function of the distribution is in the limit $n \to \infty$,

$$(1 + (u-1)q^{-k})^{I_k},$$

that is to say, the probability generating function of a binomial distribution $\mathcal{B}(I_k, q^{-k})$. As observed in [27], this asymptotic formula is even exact as soon as $n \geq kI_k$.

Theorem 8 (Knopfmacher and Knopfmacher)**.** *The probability that there are j distinct irreducible factors of degree k in a random polynomial of degree n is, for n large enough ($n \geq kI_k$), given by the binomial distribution $\mathcal{B}(I_k, q^{-k})$, namely,*

$$\Pr\{\kappa_n(k) = j\} = \binom{I_k}{j} (q^{-k})^j (1 - q^{-k})^{I_k - j}.$$

As q becomes large, the binomial probability distribution converges to a Poisson law of parameter $1/k$,

$$\Pr\{\kappa_n(k) = j\} = e^{-1/k} \frac{k^{-j}}{j!},$$

in accordance with the known distribution of cycle lengths in the random permutation model [41].

7.2. **EDF and digital trees.** We can regard the EDF phase as an abstract splitting process as follows. Start with a group G formed of k individuals. By flipping coins, separate G randomly into two subgroups, G_0 and G_1, with the probabilities for each element to be sent to G_0 and G_1 being α and β. The process is repeated recursively until all elements have been isolated. Clearly, any such recursive execution is described by a binary tree. The corresponding randomness model

$$\begin{aligned} &\Pr\left(|G_0| = k_0 \;\middle|\; |G| = k\right) = \binom{k}{k_0} \alpha^{k_0} \beta^{k-k_0}, \\ &\alpha = \frac{1}{2} - \frac{1}{2q}, \qquad \beta = \frac{1}{2} + \frac{1}{2q}, \end{aligned} \tag{32}$$

that is induced by independent splittings then coincides with the one underlying *digital trees* also known as tries [30, 35]. Amongst the many properties that are known, we mention:

— the expectation of the number of nodes in the tree (the number of splitting stages) is [30]

$$\frac{n}{H}(1 + \epsilon(n)) + o(n), \qquad H = \alpha \log_2 \frac{1}{\alpha} + \beta \log_2 \frac{1}{\beta},$$

with $\epsilon(n)$ a fluctuating function of amplitude typically $< 10^{-5}$;

— the expectation of the height of the tree is [16]

$$\frac{2}{K} \log_2 n + \mathcal{O}(1), \qquad K = \log_2(\alpha^2 + \beta^2)^{-1}.$$

Thus, these trees tend to be fairly well balanced and we expect that the cost of an EDF phase should be close to that of a perfect splitting. The lemma below provides an explicit expression.

Lemma 4. *The expected cost $C_{j,k}$ of the EDF algorithm applied to any product of j irreducible factors of degree k is*

$$\left(\frac{1}{2\alpha\beta} j(j-1) + j \sum_{m=0}^{\infty} \sum_{\ell=0}^{m} \binom{m}{\ell} \alpha^{m-\ell} \beta^{\ell} \left(1 - (1 - \alpha^{m-\ell}\beta^{\ell})^{j-1}\right) \right) (\mu_k \tau_1 + \tau_2) k^2,$$

where $\mu_k = \lambda((q^k - 1)/2) = \left\lfloor \log_2 \frac{q^k - 1}{2} \right\rfloor + \nu\left(\frac{q^k - 1}{2}\right) - 1.$

PROOF. It is convenient to regard an execution of the splitting process as a tree t and to consider, with t_0, t_1 the root subtrees, a general cost function of the additive type,

$$C[t] = e_{|t|} + C[t_0] + C[t_1]. \tag{33}$$

Here $e_{|t|}$ is a (problem specific) "toll" function that depends on the size $|t|$. that is to say, the number of irreducible factors (of degree k) to be separated.

The subtree sizes obey the Bernoulli probability of (32). Also the subproblems described by t_0, t_1 have, by design, the same characteristics as the whole tree. Thus, the expectation c_j of $C[t]$ over trees of size j satisfies the recurrence

$$c_j = e_j + \sum_{\ell=0}^{j} \binom{j}{\ell} \alpha^\ell \beta^{j-\ell} (c_\ell + c_{j-\ell}) = e_j + \sum_{\ell=0}^{j} \binom{j}{\ell} (\alpha^\ell \beta^{j-\ell} + \alpha^{j-\ell}\beta^\ell) c_\ell.$$

This translates in terms of the exponential generating functions

$$C(z) = \sum_j c_j z^j / j!, \qquad E(z) = \sum_j e_j z^j / j!,$$

into the functional equation

$$C(z) = E(z) + e^{\beta z} C(\alpha z) + e^{\alpha z} C(\beta z).$$

This difference equation iterates, leading to the explicit generating function solution

$$C(z) = \sum_{j=0}^{\infty} \sum_{\ell=0}^{j} \binom{j}{\ell} E(\alpha^{j-\ell}\beta^\ell z) e^{z(1-\alpha^{j-\ell}\beta^\ell)}. \tag{34}$$

The analysis is completed by specializing this discussion to the EDF costs. The toll function that arises from the top-level execution of the EDF procedure is then

$$\widehat{e}_j = (\mu_k \tau_1 + \tau_2)(kj)^2,$$

for $j \geq 2$. There, μ_k is the number of multiplications of the binary powering method, and the quadratic costs are induced by the naïve multiplication and gcd algorithms considered here. The toll function $e_j = j^2(1-\delta_{1,j})$ corresponds to the generating function $E(z) = z(e^z(1+z)-1)$ in (34). Extracting coefficients of the resulting generating function $C(z)$ in (34) and rescaling by $\widehat{e}_j/e_j$ then yields the statement. ∎

7.3. **Complete analysis.** Completing the analysis of EDF only requires weighting the costs given by Lemma 4 by the probability $\Pr(\kappa_n(k) = j)$ of finding j irreducible factors of degree k given by Theorem 8. By Lemma 4, the cost is of the form $\mathcal{O}(j^2k^3)$, and by Theorem 8, the probabilities are approximately $e^{-1/k}k^{-j}/j!$; thus, we expect the total cost of the DDF phase to be about

$$\sum_{k,j} (j^2k^3) \cdot \left(\frac{k^{-j}}{j!}\right) = \mathcal{O}(n^2).$$

The main result of this section gives a firm basis to this heuristic computation and determines the implied constant. In order to prove the theorem we need two technical lemmas.

Lemma 5. *When $k \to \infty$, one has for all j such that $kj \leq n$*

$$\Pr(\kappa_n(k) = j) = \frac{\binom{I_k}{j}}{q^{kj}} \left(1 + \mathcal{O}(1/k)\right),$$

where the $\mathcal{O}(1/k)$ is uniform in j. Moreover, the following uniform estimate holds

$$\Pr(\kappa_n(k) = j) = \mathcal{O}\left(\frac{1}{j!k^j}\right).$$

Proof. When $kj \leq n$, Theorem 8 yields

$$\Pr(\kappa_n(k) = j) = \frac{\binom{I_k}{j}}{q^{kj}}(1+\gamma), \quad \gamma = \sum_{\ell=1}^{N}(-1)^\ell \frac{\binom{I_k-j}{\ell}}{q^{k\ell}}, \quad N = \min([n/k]-j, I_k - j).$$

When k is large, $\gamma = \mathcal{O}(1/k)$, since

$$|\gamma| \leq \sum_{\ell=1}^{I_k-j} \frac{\binom{I_k-j}{\ell}}{q^{k\ell}} = (1+q^{-k})^{I_k-j} - 1 \leq (1+q^{-k})^{I_k} - 1 = \mathcal{O}(1/k).$$

This proves the first estimate. As for the second one, it suffices to note that

$$\binom{I_k}{j} \leq \frac{I_k^j}{j!} = \mathcal{O}\left(\frac{q^{kj}}{j!k^j}\right),$$

and to use the first estimate of the lemma. ∎

Lemma 6. *The average costs $C_{j,k}$ of Lemma 4 satisfy for all k*

$$C_{0,k} = C_{1,k} = 0, \qquad C_{2,k} = \frac{2}{\alpha\beta}\left(\mu_k \tau_1 + \tau_2\right) k^2,$$

and, uniformly,

$$C_{j,k} = \mathcal{O}(j^2 k^3).$$

Proof. The first relations are direct applications of Lemma 4. For the estimate of $C_{j,k}$, the inequality $1-(1-u)^{j-1} \leq (j-1)u$ implies

$$\begin{aligned} C_{j,k} &\leq \left(\frac{j(j-1)}{2\alpha\beta} + j\sum_{m\geq 0}\sum_{\ell=0}^{m}\binom{m}{\ell}(j-1)\alpha^{2(m-\ell)}\beta^{2\ell}\right)\left(\mu_k\tau_1+\tau_2\right)k^2 \\ &= \frac{j(j-1)}{\alpha\beta}\left(\mu_k\tau_1+\tau_2\right)k^2. \end{aligned}$$

The last equality holds since

$$\sum_{m\geq 0}\sum_{\ell=0}^{m}\binom{m}{\ell}\alpha^{2(m-\ell)}\beta^{2\ell} = \sum_{m\geq 0}(\alpha^2+\beta^2)^m = \frac{1}{2\alpha\beta}.$$

Since $\mu_k = \mathcal{O}(k)$, the statement follows. ∎

We are now ready to prove the main result of this section.

Theorem 9. *The expected cost of the EDF phase satisfies*

$$\overline{\tau EDF}_n \sim \frac{\tau_1}{\alpha\beta}\sum_{k=1}^{\lceil n/2\rceil}\mu_k, \qquad \mu_k = \left\lfloor \log_2 \frac{q^k-1}{2}\right\rfloor + \nu\left(\frac{q^k-1}{2}\right) - 1.$$

In addition, this cost is $\mathcal{O}(n^2 \log_q)$ and

$$\overline{\tau EDF}_n \sim \left(\frac{3}{4}\tau_1 \frac{q^2}{q^2-1}\log_2 q\right)(1+\xi_n)\cdot n^2, \quad -\frac{1}{3}+o(1) \leq \xi_n \leq \frac{1}{3}+o(1). \tag{35}$$

PROOF. The intuition behind the proof is that the major contribution comes from situations where just 2 factors are present, the other cases having globally a very small probability of occurrence. Let $\overline{\mathrm{E}_k}$ be the expected value of the cost of the EDF algorithm corresponding to degree k. By definition, we have $\overline{\mathrm{E}_k} = \sum_{j\geq 2} \Pr(\kappa_n(k) = j)\, C_{j,k}$, where $C_{j,k}$ is given by Lemma 4.

When $2k \leq n$, Lemma 5 and Lemma 6 entail, as $k \to \infty$,

$$\overline{EDF_k} = C_{2,k}\frac{\binom{I_k}{2}}{q^{2k}}(1+\mathcal{O}(1/k)) + \sum_{j\geq 3}\mathcal{O}\left(\frac{k^{-j}}{j!}j^2k^3\right) = \frac{\tau_1}{\alpha\beta}\mu_k + \mathcal{O}(1).$$

When $2k > n$, we have $\overline{EDF_k} = 0$. Thus, the overall cost of the EDF component is $\sum_k \overline{\mathrm{E}_k} = \frac{\tau_1}{\alpha\beta}\sum_{k=1}^{\lceil n/2\rceil}\mu_k + \mathcal{O}(n)$. The second form of the cost is obtained from the general inequality $1 \leq \nu(m) \leq 1 + \log_2 m$, upon substracting from $\nu(m)$ its "mean value" $\frac{1}{2}\log_2 m$. ∎

The quantity ξ_n in the statement measures the default of uniformity in binary representations of numbers related to the powers of q. Under the unproven assumption that such representations behave like random integers, the arithmetic function ξ_n should be close to 0. This is well supported by empirical evidence: for instance, with $q = 17$, we have

$$\xi_5 = -0.425,\ \xi_{10} = -0.060,\ \xi_{20} = -0.024,\ \xi_{50} = -0.016.$$

For all practical purposes, we may safely regard ξ_n as being asymptotic to 0.

7.4. **Equal-degree factorization in characteristic 2.** In the previous sections, we have analysed in detail the equal-degree factorization over finite fields with odd characteristic. For these cases, we have followed the algorithm by Cantor and Zassenhaus [4] who also provide a solution for the even case that relies on factoring the polynomial in a quadratic extension. Ben-Or [1] showed that this detour is not needed while proposing a method based on trace computations.

Trace computations introduce only a small change in the EDF algorithm of Fig. 4. Let m be such that $q = 2^m$. In order to compute the traces, we replace line 1 by

```
1'.    a := h + h^2 + h^(2^2) + ... + h^(2^(km-1)) mod b;
```

We observe that the analysis for the odd case is valid for the even case. First, the partitioning process is the same (with probabilities $\alpha = \beta = 1/2$). Then, the cost of computing line `1'` is the same as the cost of computing line `1` in Fig. 4. Indeed, the trace computations can be determined using basically km products of a polynomial containing j factors of degree k. This costs essentially $km(jk)^2 = k^3j^2\log q$, the same cost as in the odd case.

8. CONCLUSIONS

In this paper we have shown how the methodology of analytic combinatorics adapts to the case of polynomials over finite fields. A systematic usage of this methodology leads not only to the derivation of basic properties of random polynomials over finite fields but also to the average-case analysis of a complete polynomial factoring algorithm.

It should be clear that a large number of variants of the factorization chain can be analysed by our methods. For instance, specifics of the elimination of repeated factors stage are largely immaterial from the expected complexity standpoint since they lead to identical results in asymptotic terms. A radical possibility is then to bypass completely the first stage. In this variant, DDF not only produces the polynomials for the EDF part but also returns a polynomial

Step	Worst-case	Average-case
ERF	$\mathcal{O}(n^2)$	$\tau_2 n^2$
DDF	$\mathcal{O}(n^3 \log q)$	$0.26689\,(\lambda(q)\tau_1 + \tau_2)\, n^3$
EDF	$\mathcal{O}(n^3 \log q)$	$\left(\frac{3}{4}\tau_1 \frac{q^2}{q^2-1} \log q \cdot n^2\right)\ (1 \pm \frac{1}{3} + o(1))$

TABLE 1. A comparison of the worst cases and the average cases of the three phases of polynomial factorization.

containing the nonsquarefree part of the original polynomial. Once more, there is no difference in asymptotic terms.

Table 1 summarizes the main results of the paper in terms of the average-case analysis of the factoring algorithm and it provides a comparison with worst-case behaviors. We recall the costs of the basic operations. The cost of multiplying two polynomials of degree less than n modulo a polynomial f of degree n is $\tau_1 n^2$, and the cost of a gcd between f and a polynomial of degree less than n is $\tau_2 n^2$, for constants τ_1 and τ_2. The number of products needed to compute $h^q \bmod f$ is $\lambda(q) = \lfloor \log_2 q \rfloor + \nu(q) - 1$.

Several authors [18, 25] have stated that from a worst-case perspective, and considering fast arithmetic instead of classical one, DDF is the bottleneck for the factorization process. Our results confirm that this is also the case from an average-case perspective.

Acknowledgements. The work of Philippe Flajolet was supported by the Long Term Research Project *Alcom-IT* (# 20244) of the European Union. We are grateful to Joachim von zur Gathen for having put us in contact and for having incited us to analyse polynomial factorization in detail. Thanks also to Brigitte Vallée for several constructive suggestions regarding combinatorial decompositions.

REFERENCES

[1] BEN-OR, M. Probabilistic algorithms in finite fields. In *Proc. 22nd IEEE Symp. Foundations Computer Science* (1981), pp. 394–398.

[2] BERLEKAMP, E. *Algebraic coding theory.* McGraw Hill, New York NY, 1968.

[3] BUCHMANN, J. Complexity of algorithms in algebraic number theory. In *Number Theory. Proc. First Conf. Canadian Number Theory Assoc.* Walter de Gruyter, 1990, pp. 37–53.

[4] CANTOR, D., AND ZASSENHAUS, H. A new algorithm for factoring polynomials over finite fields. *Math. Comp. 36* (1981), 587–592.

[5] CAR, M. Factorisation dans $\mathbb{F}_q[x]$. *C. R. Acad. Sci. Paris Ser. I 294* (1982), 147–150.

[6] CARLITZ, L. The arithmetic of polynomials in a Galois field. *Amer. J. Math. 54* (1932), 39–50.

[7] CHOR, B., AND RIVEST, R. A knapsack-type public key cryptosystem based on arithmetic in finite field. *IEEE Trans. Inform. Theory. 34* (1988), 901–909.

[8] COLLINS, G. Factoring univariate integral polynomials in polynomial average time. In *Proc. EUROSAM 79* (1979), vol. 72 of *Lecture Notes in Computer Science*, pp. 317–329.

[9] COMTET, L. *Advanced Combinatorics.* Reidel, Dordrecht, 1974.

[10] DE BRUIJN, N. On the number of positive integers $\leq x$ and free of prime factors $> y$. *Indag. Math. 13* (1951), 2–12.

[11] DEDEKIND, R. Abriss einer Theorie der höhern Congruenzen in Bezug auf einen reellen Primzahlmodulus. *J. reine u. angew. Math. 54* (1857), 1–26.

[12] DICKMAN, K. On the frequency of numbers containing prime factors of a certain realtive magnitude. *Ark. Mat. Astr.Fys. 22* (1930), 1–14.

[13] FLAJOLET, P., GOURDON, X., AND PANARIO, D. Random polynomials and polynomial factorization. In *Automata, Languages, and Programming* (1996), F. Meyer auf der Heide and B. Monien, Eds., vol. 1099 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 232–243. Proceedings of the 23rd ICALP Conference, Paderborn, July 1996.

[14] FLAJOLET, P., AND ODLYZKO, A. Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics 3 2* (1990), 216–240.

[15] Flajolet, P., and Soria, M. Gaussian limiting distributions for the number of components in combinatorial structures. *Journal of Combinatorial Theory,* Series A *53* (1990), 165–182.
[16] Flajolet, P., and Steyaert, J. A branching process arising in dynamic hashing, trie searching and polynomial factorization. In *Proc. 9th ICALP Symp. 1982* (1982), vol. 140 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 239–251.
[17] von zur Gathen, J., and Panario, D. Factoring polynomials over finite fields: a survey. Tech. Rep. TR-RI-97-183, Department of Computer Science, Universität-GH Paderborn, 1997. 19 pp.
[18] von zur Gathen, J., and Shoup, V. Computing Frobenius maps and factoring polynomials. *Comput complexity 2* (1992), 187–224.
[19] Gauss, C. *Untersuchungen über Höhere Mathematik.* Chelsea, New York, 1889.
[20] Geddes, K., Czapor, S., and Labahn, G. *Algorithms for Computer Algebra.* Kluwer Academic Publishers, Boston, 1992.
[21] Goulden, I., and Jackson, D. *Combinatorial Enumeration.* John Wiley, New York, 1983.
[22] Gourdon, X. *Combinatoire, algorithmique et géométrie des polynômes.* Thèse, École Polytechnique, 1996.
[23] Gourdon, X. Largest components in random combinatorial structures. *Discrete Mathematics 180* (1998), 185–209.
[24] Greene, D., and Knuth, D. *Mathematics for the analysis of algorithms*, 3 ed. Birkhauser, Boston, 1990.
[25] Kaltofen, E., and Shoup, V. Subquadratic-time factoring of polynomials over finite fields. In *Proc. 27th ACM Symp. Theory of Computing* (1995), pp. 398–406.
[26] Knopfmacher, A. On the number of distinct degree sizes of a polynomial over a finite field. Preprint, 1996.
[27] Knopfmacher, J., and Knopfmacher, A. Counting irreducible factors of polynomials over a finite field. *SIAM Journal on Discrete Mathematics 112* (1993), 103–118.
[28] Knopfmacher, A., and Warlimont, R. Distinct degree factorizations for polynomials over a finite field. *Trans. Amer. Math. Soc. 347* (1995), 2235–2243.
[29] Knuth, D. *The art of computer programming, vol.1: fundamental algorithms*, 2 ed. Addison-Wesley, Reading MA, 1973.
[30] Knuth, D. *The art of computer programming, vol.3: sorting and searching.* Addison-Wesley, Reading MA, 1973.
[31] Knuth, D. *The art of computer programming, vol.2: seminumerical algorithms*, 2 ed. Addison-Wesley, Reading MA, 1981.
[32] Knuth, D., and Trabb-Pardo, L. Analysis of a simple factorization algorithm. *Theoretical Computer Science 3* (1976), 321–348.
[33] Lenstra, H. On the Chor-Rivest knapsack cryptosystem. *J. of Cryptology 3* (1991), 149–155.
[34] Lidl, R., and Niederreiter, H. *Finite fields*, vol. 20 of *Encyclopedia of Mathematics and its Applications.* Addison-Wesley, 1983.
[35] Mahmoud, H. *Evolution of random search trees.* John Wiley, New York, 1992.
[36] Odlyzko, A. Discrete logarithms and their cryptographic significance. In *Advances in Cryptology, Proceedings of Eurocrypt 1984* (1985), vol. 209 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 224–314.
[37] Odlyzko, A. Asymptotic enumeration methods. In *Handbook of Combinatorics*, R. Graham, M. Grötschel, and L. Lovász, Eds., vol. 2. Elsevier, 1995, pp. 1063–1229.
[38] Panario, D., Gourdon, X., and Flajolet, P. An analytic approach to smooth polynomials. Submitted, 1998.
[39] Panario, D., and Viola, A. Analysis of Rabin's polynomial irreducibility test. To appear in LATIN'98, *Latin American Theoretical INformatics*, *Lecture Notes in Computer Science*, Springer-Verlag, April 1998.
[40] Schönemann, T. Grundzüge einer allgemeinen theorie der höheren congruenzen, deren modul eine reelle primzahl ist. *J. f. d. reine u. angew. Math. 31* (1846), 269–325.
[41] Shepp, L., and Lloyd, S. Ordered cycle lengths in a random permutation. *Trans. Amer. Math. Soc. 121* (1966), 340–357.
[42] Shoup, V. On the deterministic complexity of factoring polynomials over finite fields. *Inform. Process. Lett. 33* (1990), 261–267.
[43] Shoup, V. A new polynomial factorization algorithm and its implementation. *J. Symb. Comp. 20* (1996), 363–397.
[44] Tenenbaum, G. *Introduction to analytic and probabilistic number theory.* Cambridge University Press, 1995.

P. Flajolet, Algorithms Project, INRIA Rocquencourt, F-78153 Le Chesnay, France.
E-mail address: Philippe.Flajolet@inria.fr

X.Gourdon, Algorithms Project, INRIA Rocquencourt, F-78153 Le Chesnay, France.
E-mail address: Xavier.Gourdon@inria.fr

D. Panario, Dept. of Computer Science, University of Toronto, Toronto, Canada M5S-3G4.
E-mail address: daniel@cs.toronto.edu

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105,
78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS
Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
(France)
http://www.inria.fr
ISSN 0249-6399