CAMBRIDGE TRACTS IN MATHEMATICS

General Editors

B. BOLLOBÁS, W. FULTON, A. KATOK, F. KIRWAN,
P. SARNAK, B. SIMON, B. TOTARO

175 The Large Sieve and its Applications:
Arithmetic Geometry, Random Walks and Discrete Groups

The Large Sieve and its Applications

Arithmetic Geometry, Random Walks and Discrete Groups

E. KOWALSKI
Swiss Federal Institute of Technology (ETH), Zürich

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521888516

First published 2008

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-88851-6 hardback

Pour les soixante ans de Jean–Marc Deshouillers

Contents

Preface

> 'The Romans,' Roger and the Reverend Dr. Paul de la Nuit were drunk together one night, or the vicar was, 'the ancient Roman priests laid a sieve in the road, and then waited to see which stalks of grass would come up through the holes.'
>
> *Thomas Pynchon, 'Gravity's Rainbow'*

These notes arose, by the long and convoluted process that research often turns out to be, from a supposedly short addition to my paper [**80**]. This is a story that is certainly typical of much of scientific research, and since I always find this fascinating, and hardly visible from the outside once a paper or book is published,[1] I will summarize the events briefly. Readers who like science rather dry or dour may wish to start reading Chapter 1.

The original ambition was simply to extend the large sieve bound for Frobenius conjugacy classes of this first paper to the stronger form classically due to Montgomery, which would mean that 'small sieve' applications would become possible. The possibility of this extension seemed clear to me, as well as the relative paucity of new applications.[2] At the same time, it seemed natural to 'axiomatize' the setting in a way allowing an identical treatment of the classical large sieve inequality and this newer variant, and this seemed a worthwhile enough goal.

All this should not have taken very long, either in time or space, except that inevitable delays due to teaching and other duties led to the thought that maybe other applications of this abstract form of sieve would be possible, and could be

[1] A striking recent instance of this process is described by A. Wiles in the introduction to his paper proving Fermat's Great Theorem.

[2] In large sieve situations, applying the best small sieve bound gives very small gains, whereas small sieve cases, by definition, can be handled by small sieves, which were already sufficiently general to handle the 'obvious' applications, and in fact strong enough to prove lower bounds in some contexts.

briefly discussed in the course of the paper, which would thus become stronger. A natural fit, given my background and the emphasis on random matrices as an interpretation of the results of [**80**], was to think of trying to prove, e.g., that a 'generic' unimodular integral $n \times n$ matrix has irreducible characteristic polynomial (or maximal splitting field), as an application of the large sieve applied to $SL(n, \mathbf{Z})$. I started thinking about this problem, seeing clearly that harmonic analysis of automorphic forms on $SL(n, \mathbf{Z})\backslash SL(n, \mathbf{R})$ would be called for, and that this would require some learning on my part for $n \geqslant 3$. Clearly this would be material for *another* paper, a quite interesting one since I knew of no previous use of sieve in such situations. Because of the strong link to spectral theory of automorphic forms, I was pretty sure I would have heard of it if published papers on this topic existed; as it was, there were results of Duke, Rudnick and Sarnak [**33**] (and their later extensions) giving asymptotic formulas for the *number* of unimodular matrices with bounded norm, but not for the more general 'exponential sums' arising from the sieve theory.

In the meantime, D. Zywina sent me his preprint ('*The large sieve and Galois representations*', 2007) which contained a slightly different formulation of an abstract form of the large sieve, with applications to distribution of Frobenius elements in number fields, specifically to the Lang–Trotter Conjecture. His sieve axioms were in many respects more general than mine, except for one condition which I had to introduce in [**80**] because of specific features of the arithmetic of varieties over finite fields (the difference between arithmetic and geometric fundamental groups). Still, where his conditions were more general, I could in fact very easily assume the same generality, and reading his preprint led me to rewrite mine in this light. This did not bring new applications. On the other hand, as I was reading (mostly for the pleasure of it) the nice book by P. de la Harpe on geometric group theory [**57**], I thought that one could also try to use as targets of sieves the subsets of groups defined by word length (with respect to some system of generators) being smaller than some quantity. However, not knowing much about this topic, this was mostly speculative.

But around the same time, I. Rivin posted a preprint [**108**] on arXiv (`www.arXiv.org`) which directly mentioned the problem of irreducibility of characteristic polynomials of unimodular matrices. He also mentioned the results of Duke, Rudnick and Sarnak but did not prove that 'most' matrices have this property. What he managed to prove was an analogue of the more combinatorial variant: instead of looking at balls in the word-length metric, rather he was looking at random walks on the group of length $k \to +\infty$. His method for detecting irreducibility was similar to the 'old' method used by van der Waerden for integral polynomials with bounded height, combined with results of Chavdarov [**22**] (which already played a role in [**80**], one of the results

of which was indeed a strong quantitative strengthening of Chavdarov's main result, following Gallagher's large sieve strengthening [**46**] of van der Waerden's result), and in particular the statement proved was qualitative and did not give explicit bounds for the probability of having a reducible characteristic polynomial.

A remarkable novel feature of Rivin's work was the new applications he discussed, which concerned 'generic' properties of automorphisms either of compact connected surfaces or free groups. In each case, the action of such elements on a free abelian group (the homology of the surface or abelianization of the free group, respectively) was sufficient to detect an interesting condition by looking at the corresponding characteristic polynomial. Rivin thus proved in a very simple way a (special case of a) result of Maher [**96**]: the probability that the k-th step of a random walk on the mapping class group of a surface of genus g is pseudo-Anosov tends to 1 as $k \to +\infty$.

As I mentioned to Rivin that I had been working with the large sieve with applications to characteristic polynomials in mind, he told me that Bourgain, Gamburd and Sarnak were investigating issues related to sieve in arithmetic groups and forwarded their preprint [**14**]. This work was, in small sieve contexts, concerned with showing that orbits of certain subgroups G of arithmetic groups acting on $\mathbf{Z}^n$ contain infinitely many points with prime (or almost prime) coordinates. What was clearly explained was that, apart from fairly standard sieve machinery going back to Brun or Selberg, the crucial feature that must be exploited (and proved) is the expanding property of congruence quotients of the group G.

As I became aware of these very interesting developments, my paper remained unchanged. Or rather, what was expanding in it was a 'sidebar' having to do with natural questions suggested by the sieve framework: what is the largest dimension of an irreducible representation of a finite group of Lie type, such as $SL(n, \mathbf{Z}/\ell\mathbf{Z})$ or $Sp(2g, \mathbf{Z}/\ell\mathbf{Z})$, and what is the sum of those dimensions? This had already puzzled me while writing [**80**], where I used 'trivial' bounds for those quantities. As I tried once more to get some understanding of the theory of Deligne–Lusztig characters which describes the representations of such groups, I finally wrote to F. Digne and J. Michel, with the feeling that this must certainly be known, but hidden somewhere inaccessible to 'simple' searches in *Mathematical Reviews*. However, J. Michel did not know if the first question had been considered (he pointed out the papers of Gow [**50**] and Vinroot [**129**] concerning the second problem). Based on his indications, I managed to write down a proof of the estimate which I had found 'reasonable' to expect.

Finally summer vacation came. Then, in a short time, I found and wrote down a new amusing application of the sieve to the study of denominators of rational points on elliptic curves, which was a good example of the 'abstract' framework. More importantly, Rivin's use of random walks prompted me to generalize the sieve context to that of estimating the measure of some 'sifted set', and not necessarily its cardinality, in order to incorporate applications having to do with general random walks. And using Property (τ) for discrete groups together with some nice probabilistic ideas described in the survey on random walks on groups by L. Saloff-Coste [**111**], I obtained an effective form of Rivin's irreducibility theorem for random walks on $SL(n, \mathbf{Z})$ or $Sp(2g, \mathbf{Z})$.

At this point, I felt that I merely needed to polish a few things and then send the paper to a well-chosen journal. I was wondering if splitting it into multiple parts might not be better (something I usually strongly dislike), since its growing mathematical spread, while appealing, obviously made it difficult to find a single referee: by this time, the crucial insights were from analytic number theory, the tools ranged from representation theory, including Deligne–Lusztig theory, to Property (τ) and the Riemann Hypothesis over finite fields, not to mention the use of probabilistic vocabulary. And familiarity with [**80**] was quite obviously assumed ...

But then I realized that the very basic formal part of the large sieve was unduly complicated and framed in the wrong way, bonding the method with group theory much too early (the title at the time was 'The algebraic principle of the large sieve', a joking pun on [**98**]). By moving the group theory to a different part of the argument (the choice of a suitable orthonormal basis for finite-dimensional Hilbert spaces), the principle of the sieve could be both simplified and generalized once more. In retrospect, nothing seems more obvious, but the simpler form had been completely obscured by the force of habit together with the fact that all applications I knew were linked with a group and its representation theory.

So I rewrote much of the beginning part and adapted the rest; by this time the paper was around 55 (full) pages long. After some more hesitation, some more feature-creep, and taking advice from P. Sarnak and A. Granville, getting this text in a journal seemed less and less practical. Because of the many applications, I wanted the paper to be accessible to as large an audience as possible, and the style of the writing appeared to me to become unsuitable for, say, geometers interested in the stronger form of Maher's and Rivin's results (I had realized, looking at [**96**] quite late, that my bound for characteristic polynomials of $Sp(2g, \mathbf{Z})$ implied a solution to a further question of Maher, namely the *transience* of the set of non-pseudo-Anosov elements during a random walk on the mapping class group).

The outcome of this process is that I have expanded the paper to a short book, adding brief surveys of most of the important material that may not be known to all readers. This includes the representation theory of finite groups, Property (τ) (and Property (T)) – with a sketch of the proof of Property (T) for $SL(n, \mathbf{Z})$ due to Shalom [124], sums of multiplicative functions, probability theory and random walks, and the mapping class groups of surfaces. Of course, for some of these, I have no claim to expertise and the surveys should only be thought of as delineating the basic definitions and some basic information which I found especially interesting (or beautiful!) while learning about the subject.

All this will, I hope, have both the effect of making the text readable for non-analytic number theorists that may have potential use of ideas related to the large sieve, and to make analytic number theorists aware of some potential areas where their ideas might be useful.

Acknowledgments

As the preface shows, a lot of people have had a great influence on the final appearance of this work beyond the impetus of [**80**]. I mention again in particular D. Zywina, who developed an abstract setup of the large sieve similar to the conjugacy sieve described in Chapter 3, which prompted me on more than one occasion to evolve my own version; and I. Rivin, whose work suggested the probabilistic sieve setting, and who also mentioned to me the work of Bourgain, Sarnak and Gamburd. I also wish to thank P. Sarnak for sending me a copy of his email to his coauthors. Finally, I thank J. Michel for providing the ideas of the proof of Proposition 5.5 and explaining some basic properties of representations of finite groups of Lie type; M. Burger for information concerning Property (T) and Property (τ), and for correcting my misunderstanding of his paper [**18**]; P. Duchon and M.-L. Chabanol for help, advice and references concerning probability theory and graph theory; K. Belabas for suggestions concerning numerical experiments; D. Khoshnevisan for providing a correct proof of one probabilistic statement; and J. Wu for explaining some points concerning [**88**]. Also, I wish to thank F. Jouve for finding many small mistakes and imprecisions in the original drafts.

Work on this book was partially supported by the ANR (L'Agence Nationale de la Recherche) Project ARITHMATRICS. Some preliminary results were presented during the conference organized by this project in Bordeaux in April 2006, and the remarks of participants were very helpful in shaping the later evolution of the ideas presented here. A much shorter preliminary version of this book was also posted on arXiv as `arXiv:math.NT/0610021`.

Prerequisites and notation

There are two types of readers for whom this book is written: some who are knowledgeable about analytic number theory, and maybe very familiar with sieve methods, and who (we hope) will find the new and unfamiliar applications of interest; and some who are interested in a specific application (e.g., those around properties of mapping class groups, or zeta functions of algebraic varieties over finite fields, or random walks on discrete groups), but not necessarily in all of them, and who may not be familiar with the principles of analytic number theory.

Fortunately, there is in fact very little prerequisite for most of the book; the basic principle of the large sieve uses nothing more than basic linear algebra and analysis (finite-dimensional Hilbert spaces). When it comes to applications, where more sophisticated tools are often involved, we follow the policy of defining from scratch all notions that appear, and provide the reader with precise references for all facts we use about such topics as elliptic curves, discrete groups, algebraic groups, random walks and harmonic analysis. The only (partial) exception is in Chapter 8 where we need the machinery of ℓ-adic sheaves over finite fields, and their cohomology. But even then, the statements of the applications of the sieve (at least) should be understandable by any reader, and we hope that the mechanism of the proofs is explained clearly enough that analytic number theorists will be able to benefit from reading this chapter.

We now summarize the most common notation. Less standard notation will be explained in each chapter when first used (see in particular the beginning of Chapter 2), and moreover the appendices contain quick surveys of the definitions of (almost) all mathematical terms which occur in the book.

As usual, $|X|$ denotes the cardinality of a set; however if X is a measure space with measure μ, we sometimes write $|X|$ instead of $\mu(X)$.

By $f \ll g$ for $x \in X$, or $f = O(g)$ for $x \in X$, where X is an arbitrary set on which f is defined, we mean synonymously that there exists a constant

$C \geqslant 0$ such that $|f(x)| \leqslant Cg(x)$ for all $x \in X$. The 'implied constant' is any admissible value of C. It may depend on the set X which is always specified or clear in context. The notation $f \asymp g$ means $f \ll g$ and $g \ll f$. On the other hand $f(x) = o(g(x))$ as $x \to x_0$ is a topological statement meaning that $f(x)/g(x) \to 0$ as $x \to x_0$. We also use the $O()$ notation in other types of expressions; the meaning should be clear: e.g., $f(x) \leqslant g(x) + O(h(x))$ for $x \in X$, means that $f \leqslant g + h_1$ in X for some (non-negative) function h_1 such that $h_1 = O(h)$. (For instance, $x \leqslant x^2 + O(1)$ for $x \geqslant 1$, but it is not true that $x - x^2 = O(1)$.)

In this book, any statement of a lemma, proposition, theorem or corollary will include an explicit mention of which parameters the 'implied constant' depends on; any divergence from this principle is an error, and the author should be made aware of it. The same explicitness will be true for many, but not all, of the intermediate statements (where sometimes it will be clear enough what the parameters involved are, from the flow of the argument). This insistence may look pedantic, but uniformity in parameters is crucial to many applications of analytic number theory, and this should make the text usable by all mathematicians with confidence that there is no hidden dependency. (Algebraic-minded readers may note that indicating the dependency of those parameters is somewhat analogous to stating explicitly in which category a morphism between two objects is defined; the author's experience is that not having this information clearly stated *even if it is completely obvious for knowledgeable readers* can create a lot of confusion for beginners.)

For a group G, $G^\sharp$ denotes the set of its conjugacy classes, and for a conjugacy-invariant subset $X \subset G$, $X^\sharp \subset G^\sharp$ is the corresponding set of conjugacy classes. The conjugacy class of $g \in G$ is denoted $g^\sharp$.

For q a power of a prime number, $\mathbf{F}_q$ denotes a finite field with q elements.

Unless otherwise specified (as in Chapter 5), p always denotes a prime number. If $n \geqslant 1$ is an integer, sums or products over divisors of n always mean divisors $d \geqslant 1$. We use standard arithmetic functions φ, ψ, ω and μ,[3] defined as follows for an integer $n \geqslant 1$ in terms of the prime factors of n:

$$\varphi(n) = n \prod_{p|n} (1 - p^{-1}), \quad \psi(n) = n \prod_{p|n} (1 + p^{-1}), \quad \omega(n) = |\{p \mid p \mid n\}|,$$

$$\mu(n) = \begin{cases} (-1)^k & \text{if } n = p_1 \cdots p_k \text{ with } p_1 < \cdots < p_k, \\ 0 & \text{otherwise,} \end{cases}$$

[3] No confusion should arise with measures also denoted μ.

We denote as (a, b) the greatest common divisor of integers a and b, unless this creates ambiguity with pairs of integers. Similarly, $[a, b]$ is the least common multiple. An integer $n \geqslant 1$ is *squarefree* if it is not divisible by the square of a prime p, or equivalently if $\mu(n) \neq 0$. We use the shorthand notation

$$\sum_m^\flat \alpha(m)$$

for a sum restricted to squarefree integers m.

We denote by $\pi(x)$ the prime counting function, i.e., the number of primes $p \leqslant x$, and by $\pi(x; q, a)$ the prime counting function in arithmetic progressions, i.e., the number of primes $p \leqslant x$ which are congruent to a modulo q. Of course, $\pi(x; q, a)$ is bounded if and only if $(a, q) \neq 1$ (by Dirichlet's theorem on primes in arithmetic progressions).

We recall some asymptotic formulas of prime number theory, the second of which is a strong form of the Prime Number Theorem:

$$\sum_{p \leqslant x} \frac{1}{p} = \log\log x + O(1), \qquad \pi(x) = \frac{x}{\log x} + O\Big(\frac{x}{(\log x)^2}\Big),$$

for $x \geqslant 3$.

For $z \in \mathbf{C}$, we denote $e(z) = \exp(2i\pi z)$, so that $e(\cdot)$ is a non-trivial homomorphism $\mathbf{C}/\mathbf{Z} \to \mathbf{C}^\times$.

In probabilistic contexts, $\mathbf{P}(A)$ is the probability of an event, $\mathbf{E}(X)$ is the expectation of a random variable X, $\mathbf{V}(X)$ its variance, and $\mathbf{1}_A$ is the characteristic function of an event A. See Appendix F for the basic definitions.

Let k be a field, and V a k-vector space of even dimension $\dim V = 2g$. If $\langle \cdot, \cdot \rangle : V \times V \to k$ is a non-degenerate alternating bilinear form on V, we denote by $Sp(V)$, $Sp(\langle \cdot, \cdot \rangle)$ or more commonly by $Sp(2g, k)$ the *symplectic group* of V, namely the group of invertible linear transformations of V preserving this bilinear form; it is the group of those $g \in GL(V)$ such that

$$\langle gv, gw \rangle = \langle v, w \rangle$$

for all $v, w \in V$. The notation $Sp(2g, k)$ is justified by the fact that, up to isomorphism, there is only one non-degenerate alternating bilinear form on V. If a specific model is needed, one can fix a vector space W of dimension g, and put $V = W \oplus W'$, where W' is the dual of W, and let

$$\langle (v_1, \ell_1), (v_2, \ell_2) \rangle = \ell_1(v_2) - \ell_2(v_1).$$

The subspaces W and W' are then instances of *Lagrangian subspaces*, i.e., subspaces of maximal dimension g such that the restriction of the alternating form to the subspace is identically zero. All Lagrangian subspaces of V are

images of any fixed one (such as W above) by an element of $Sp(V)$, i.e., $Sp(V)$ acts transitively on the set of Lagrangian subspaces. If W_1, W_2 are Lagrangian subspaces, they are *transverse* if $W_1 \cap W_2 = 0$, or equivalently if both together span V.

Moreover, we denote by $CSp(V)$, $CSp(\langle \cdot, \cdot \rangle)$ or $CSp(2g, k)$ the group of *symplectic similitudes*, i.e., of those $g \in GL(V)$ such that

$$\langle gv, gw \rangle = m(g)\langle v, w \rangle$$

for all $v, w \in V$, where $m(g) \in k^\times$ is a scalar called the *multiplicator* of g. This is a surjective group homomorphism, and there is therefore an exact sequence

$$1 \to Sp(V) \to CSp(V) \xrightarrow{m} k^\times \to 1.$$

We recall the formulas for the cardinality of $GL(n, \mathbf{F}_q)$ and $Sp(2g, \mathbf{F}_q)$ for a finite field $\mathbf{F}_q$ with q elements:

$$|GL(n, \mathbf{F}_q)| = \prod_{k=0}^{n-1} (q^n - q^k) = q^{n(n-1)/2} \prod_{k=1}^{n} (q^k - 1), \tag{0.1}$$

$$|Sp(2g, \mathbf{F}_q)| = q^{g^2} \prod_{k=1}^{g} (q^{2k} - 1). \tag{0.2}$$

When working with matrices $g \in M(n, A)$, where A is a commutative ring with unit, we will consider both the standard *characteristic polynomial* of g, namely $\det(T - g) \in A[T]$, which is a monic polynomial of degree n taking value $(-1)^n \det(g)$ at 0; and the *reversed characteristic polynomial* $\det(\mathrm{Id} - Tg) \in A[T]$, where Id is the identity matrix. This is of degree equal to the rank of g, takes value 1 at 0, and has leading term $\det(g)T^n$ if g is invertible. Obviously, whenever invertible matrices are considered, all results on either of these can be restated in terms of the other, or of $\det(g - T)$: we have

$$\det(\mathrm{Id} - gT) = T^n \det(T^{-1} - g).$$

If we wish to speak of the characteristic polynomial of an endomorphism of a free A-module V of finite rank, we write $\det(T - A \mid V)$ or $\det(\mathrm{Id} - TA \mid V)$.

If G is a group, $[G, G]$ is the commutator subgroup, generated by commutators $[x, y] = xyx^{-1}y^{-1}$ for $x, y \in G$, and the abelian group $G/[G, G]$ is the *abelianization* of G.

The symmetric group on n letters is denoted $\mathfrak{S}_n$. Moreover, for $g \geqslant 1$, W_{2g} denotes the group of *signed permutations* of g pairs $(2i - 1, 2i)$, $1 \leqslant i \leqslant 2g$,

i.e., the subgroup of elements $\sigma \in \mathfrak{S}_{2g}$ such that $\sigma(\{2i-1, 2i\})$ is a pair $\{2j, 2j-1\}$ for all i. This group has order $2^g g!$ and sits in an exact sequence

$$1 \to \{\pm 1\}^g \to W_{2g} \xrightarrow{p} \mathfrak{S}_g \to 1,$$

where the right-hand map assigns to $\sigma \in W_{2g}$ the permutation of the g pairs $(2i-1, 2i)$, the natural generators σ_i of the kernel being the signed permutations which act as the identity except for $\sigma(2i-1) = 2i, \sigma(2i) = 2i-1$.