Annals of Mathematics **175** (2012), 1507–1550
http://dx.doi.org/10.4007/annals.2012.175.3.10

On sharp transitions in making squares

By Ernie Croot, Andrew Granville, Robin Pemantle, and Prasad Tetali

Abstract

In the fastest-performing integer factoring algorithms, one creates a sequence of integers (in a pseudo-random way) and wishes to rapidly determine a subsequence whose product is a square. In 1994 Pomerance stated the following problem which encapsulates all of the key issues: *Select integers $a_1, a_2, \dots$, at random from the interval $[1, x]$, until some (nonempty) subsequence has product equal to a square. Find a good estimate for the expected stopping time of this process.* A good solution should allow one to determine the optimal choice of parameters in many factoring algorithms.

Pomerance (1994), using an idea of Schroeppel (1985), showed that with probability $1 - o(1)$ the first subsequence whose product equals a square occurs after at least $J_0^{1-o(1)}$ integers have been selected, but no more than J_0, for an appropriate (explicitly determined) $J_0 = J_0(x)$. We tighten Pomerance's interval to

$$[(\pi/4)(e^{-\gamma} - o(1))J_0,\ (e^{-\gamma} + o(1))J_0],$$

where $\gamma = 0.577...$ is the Euler-Mascheroni constant, and believe that the correct interval is $[(e^{-\gamma} - o(1))J_0,\ (e^{-\gamma} + o(1))J_0]$, a "sharp threshold". In our proof we confirm the well-established belief that, typically, none of the integers in the square product have large prime factors.

The heart of the proof of our upper bound lies in delicate calculations in probabilistic graph theory, supported by comparative estimates on smooth numbers using precise information on saddle points.

1. Introduction

Several algorithms for factoring integers n (including Dixon's random squares algorithm [6], the quadratic sieve [10], the multiple polynomial quadratic sieve [14], and the number field sieve [2] — see [13] for a nice expository article on factoring algorithms) work by generating a pseudo-random sequence of integers $a_1, a_2, \dots$, with each

$$a_i \ \equiv \ b_i^2 \pmod{n}$$

E.C.: Supported in part by an NSF and by an NSA grant. A.G.: Partiellement soutenu par une bourse de la Conseil de recherches en sciences naturelles et en génie du Canada. R.P.: Supported in part by NSF Grant DMS-01-03635. P.T.: Supported in part by NSF Grants DMS-0401239 and DMS-0701043.

until some subsequence of the a_i's has product equal to a square. Say we have such a subsequence

$$a_{i_1}, \dots, a_{i_k}, \text{ where } Y^2 = a_{i_1} \cdots a_{i_k}$$

and set

$$X^2 = (b_{i_1} \cdots b_{i_k})^2.$$

Then

$$n \mid Y^2 - X^2 = (Y - X)(Y + X),$$

and there is a fair chance that $\gcd(n, Y - X)$ is a nontrivial factor of n. If so, we have factored n.

In his lecture at the 1994 International Congress of Mathematicians, Pomerance [11], [12] observed that in the (heuristic) analysis of such factoring algorithms one assumes that the pseudo-random sequence $a_1, a_2, \dots$ is close enough to random that we can make predictions based on this assumption. Hence it makes sense to formulate this question in its own right, in particular to determine whether this part of the factoring algorithm can be significantly sped up.

Pomerance's Problem. Select positive integers $a_1, a_2, \dots \le x$ independently at random (that is, $a_j = m$ with probability $1/x$ for each integer m, $1 \le m \le x$) until some subsequence of the a_i's has product equal to a square. When this occurs, we say that the sequence has a *square dependence*. What is the expected stopping time of this process?

To discuss the history of this problem, and our own work, we need to introduce some notation. Let $\pi(y)$ denote the number of primes up to y. Call n a *y-smooth integer* if all of its prime factors are $\le y$, and let $\Psi(x, y)$ denote the number of y-smooth integers up to x. Let $y_0 = y_0(x)$ be a value of y which maximizes $\Psi(x, y)/y$, and let

$$J_0(x) := \frac{\pi(y_0)}{\Psi(x, y_0)} \cdot x. \tag{1}$$

(We will see later, in (9), that $\log J_0(x) \sim 2 \log y_0(x) \sim \sqrt{2 \log x \log\log x}$.) In Pomerance's problem, let T be the smallest integer t for which $a_1, \dots, a_t$ has a square dependence. (Note that T is itself a random variable.) In 1985, Schroeppel gave a simple argument to justify that for any $\varepsilon > 0$, we have

$$\text{Prob}(T < (1 + \varepsilon) J_0(x)) = 1 - o(1)$$

as $x \to \infty$, and in 1994 Pomerance showed that

$$\text{Prob}(T > J_0(x)^{1-\varepsilon}) = 1 - o(1)$$

as $x \to \infty$. Therefore there is a transition from "unlikely to have a square product" to "almost certain to have a square product" at $T = J_0(x)^{1+o(1)}$.

An estimate with an exponent of $1+o(1)$ is actually quite weak. To address this problem, Pomerance asked in [12] whether there is a sharper transition. We conjecture that T has a *sharp threshold*: This would mean that there exists a function $f(x)$ such that for every $\varepsilon > 0$,

$$\text{Prob}(T \in [(1-\varepsilon)f(x),\ (1+\varepsilon)f(x)]) \ = \ 1 - o(1) \tag{2}$$

as $x \to \infty$. In fact, we believe that this threshold is $f(x) = e^{-\gamma}J_0(x)$.

CONJECTURE 1.1. *For every $\varepsilon > 0$, we have*

$$\text{Prob}(T \in [(e^{-\gamma}-\varepsilon)J_0(x),\ (e^{-\gamma}+\varepsilon)J_0(x)]) \ = \ 1 - o(1) \tag{3}$$

as $x \to \infty$, where $\gamma = 0.577\cdots$ is the Euler-Mascheroni constant.

The constant $e^{-\gamma}$ in this conjecture is well known to number theorists. It appears as the ratio of the proportion of integers free of prime divisors smaller than y, to the proportion of integers up to y that are prime. However, this is not how it appears in our discussion, and we have failed to find a more direct route to this prediction.

The bulk of this article will be devoted to establishing the upper bound in the above conjecture. We will prove something a little weaker than the conjectured lower bound.

THEOREM 1.2. *We have*

$$\text{Prob}(T \ \in \ [(\pi/4)(e^{-\gamma}-\varepsilon)J_0(x),\ (e^{-\gamma}+\varepsilon)J_0(x)]) \ = \ 1 - o(1)$$

for any $\varepsilon > 0$ as $x \to \infty$.

To obtain the lower bound in our theorem, we used a "first moment method" approach, which is not straightforward for the following reason. The methods of Pomerance and Schroeppel lead to a sequence of more precise estimates for the expectation of T by considering more terms in the expansion of $\log T$. This leads, however, to an unwieldy infinite sum, whose limit does not appear tractable. Expresssing T instead in terms of saddle points as in the work of Hildebrand and Tennenbaum [9] allowed us to bypass this problem.

Schroeppel established his upper bound, $T \le (1+o(1))J_0(x)$, by showing that by then one expects more than $\pi(y_0)$ y_0-smooth integers amongst $a_1, a_2, \ldots, a_T$, which guarantees that the sequence has a square dependence. (To see this, create a matrix over $\mathbb{F}_2$ whose columns are indexed by the primes up to y_0, whose rows are indexed by the numbers i such that a_i is y_0-smooth, and whose (i,p)th entry is given by the exponent on p in the factorization of a_i, for each y_0-smooth a_i. Then a square dependence amongst the a_i is equivalent to a dependence amongst the corresponding rows of our matrix so that we are guaranteed a square dependence once the matrix has more than $\pi(y_0)$ rows.)

If we replace the complicated random model that creates this matrix by one in which any given row appears as a row of this matrix with equal probability, then one expects a linear dependence only once the matrix has more than $\pi(y_0) - O(1)$ rows. (See [5, §3.1] for details; also see [3] for a lower bound in a related model of choosing binary vectors of fixed weight randomly, until finding a $GF(2)$-dependent set.)

Schroeppel's approach is not only good for theoretical analysis; in practice one searches among the a_i for y_0-smooth integers and hunts amongst these for a square dependence, using linear algebra in $\mathbb{F}_2$ on the primes' exponents. Computing specialists have also found that it is easy and profitable to keep track of a_i of the form $s_i q_i$, where s_i is y_0-smooth and q_i is a prime exceeding y_0; if both a_i and a_j have exactly the same large prime factor $q_i = q_j$, then their product is a y_0-smooth integer times a square and so can be used in our matrix as an extra smooth number. This is called the *large prime variation*, and the upper bound in Theorem 1 of [5] is obtained by computing the limit of this method (to obtain a constant, in place of $e^{-\gamma}$ which is a tiny bit smaller than $3/4$).

One can also consider the *double large prime variation* in which one allows two largish prime factors so that, for example, the product of three a_is of the form pqs_1, prs_2, qrs_3 can be used as an extra smooth number. Experience has shown that each of these variations has allowed a small speed up of various factoring algorithms (though at the cost of some nontrivial extra programming), and a long open question has been to formulate all of the possibilities for multi-large prime variations and to analyze how they affect the running time. Sorting out this combinatorial maze has been the most difficult part of our work.

When our process terminates (at time T) we have some subset I of $a_1, \dots, a_T$, including a_T, whose product equals a square.[1] It is not hard to show that this square product is T^2-smooth with probability $1 - o(1)$ (see [5, §3.2]); here we give a more precise idea of what I looks like.

THEOREM 1.3. (a) *In the special case that for $\varepsilon > 0$, conditional on the event $\{T < (\pi/4)(e^{-\gamma} - \varepsilon)J_0(x)\}$, we find that I consists of a single number a_i (which is therefore a square) with probability $1 - o(1)$.*

(b) *In general, with probability $1 - o(1)$, we have that*

$$y_0 \exp(-(c_3 + \varepsilon)\sqrt{\log y_0}) \ \le \ |I| \ \le \ y_0 \exp((c_3 + \varepsilon)\sqrt{\log y_0})], \tag{4}$$

[1]Note that I is unique, else if we have two such subsets I and J, then $(I \cup J) \setminus (I \cap J)$ is also a set whose product equals a square but does not contain a_T, and so the process would have stopped earlier than at time T.

where $c_3 = \sqrt{2 - \log 2}$. *In other words, when the algorithm terminates the square product* I *is, almost certainly, composed of* $y_0^{1+o(1)} = J_0(x)^{1/2+o(1)}$ *numbers* a_i.

(c) *Also, with probability* $1 - o(1)$, *all the elements of* I *are*

$$y_0^2 \exp((2+\varepsilon)\sqrt{\log y_0 \log\log y_0})\text{−smooth.}$$

The last part of this result confirms the long held suspicion that the earliest occurring square products are almost always composed only of smooth numbers with a suitable smoothness parameter, though the smoothness bound that we give may be significantly larger than is possible, for all we know.

We expect that one can give more precise descriptions of I, specifying more precisely how large I is and improving the smoothness bound on the elements of I, perhaps even to $y_0\phi(x)$ for any function ϕ for which $\phi(x) \to \infty$ as $x \to \infty$.

There are now several theorems along the lines of Conjecture 1.1 in the literature, including some quite general approaches. Friedgut's theorem [8], characterizing a *coarse threshold* for monotone or symmetric[2] graph properties, has been instrumental in proving the existence of a sharp threshold for several graph properties. However it does not seem to be applicable in the present context since the square dependence problem is not symmetric. Bourgain's strengthening of sorts of Friedgut's theorem (see the appendix to [8]) is in principle applicable in the present context, though various researchers have not yet succeeded in doing so.

Pomerance's main goal in enunciating the random squares problem was to provide a model that would prove useful in analyzing the running time of factoring algorithms, such as the quadratic sieve. In [5] we analyzed the running time of Pomerance's random squares problem to show that the running time will be inevitably dominated by finding the actual square product once we have enough integers. Indeed this carries over to an analysis of the quadratic sieve factoring algorithm (and presumably the other factoring algorithms as well); a consequence is that to optimize the running time of the quadratic sieve we look for a square dependence among the y-smooth integers with y significantly smaller than y_0, so that Pomerance's problem is not quite so germane to the question as it had at first appeared. See [5] for further discussion of these issues.

In discussion, David Moulton noted that a slight variation of Pomerance's problem allows us to fully analyze a slight variation of Dixon's random squares algorithm; we will give details at the end of Section 5.

[2]That is, invariant under permutations of the elements involved.

The paper is organized as follows. In Section 2, we derive the necessary technical lemmas involving smooth numbers. In Section 3, we derive the lower bound for T given in Theorem 1.2 and develop these ideas to prove Theorem 1.3. In Section 4, we develop our analysis of multiprime variations. Finally, in Section 5, we discuss the actual implications for factoring algorithms of our results.

Acknowledgements. We wish to thank David Moulton for allowing us to include his argument reducing Dixon's original algorithm to Pomerance's problem; thanks also to a referee for suggestions that led to streamlining the proof of the upper bound.

2. **Smooth numbers**

In previous analyses of these questions, authors have typically used estimates for $\Psi(x,y)$ for y a fixed power of y_0. In this range one can determine an asymptotic for $\Psi(x,y)$ in terms of a saddle point, an implicit quantity. It has proved to be difficult to deduce an asymptotic for $\Psi(x,y)$, or even something close, in terms of simple explicit functions. One of the key innovations in this article is to by-pass this issue by comparing values of $\Psi(x,y)$ for different, but closely related, values of x and y. Since the saddle points are not too different, one can obtain sharp explicit estimates for the ratio of two such Ψ-values. In this technical section we deduce several such results, primarily from the deep work of Hildebrand and Tenenbaum [9], which will be useful later.

2.1. *Classical smooth number estimates.* From [9] we have that the estimate

$$\Psi(x,y) = x\rho(u)\left\{1 + O\left(\frac{\log(u+1)}{\log y}\right)\right\}, \quad \text{where} \quad x = y^u \tag{5}$$

holds in the range

$$\exp\left((\log\log x)^2\right) \ \le \ y \ \le \ x, \tag{6}$$

where $\rho(u) = 1$ for $0 \le u \le 1$, and where

$$\rho(u) = \frac{1}{u}\int_{u-1}^{u} \rho(t)\,dt \ \text{ for all } u > 1.$$

This function $\rho(u)$ satisfies

$$\rho(u) \ = \ \exp(-u(\log u + \log\log u - 1 + o(1))) \ = \ \exp(-(u+o(u))\log u);$$

and so

$$\Psi(x,y) \ = \ x\left(\frac{e+o(1)}{u\log u}\right)^u \ = \ x/u^{u+o(u)}. \tag{7}$$

Now let

$$L := L(x) = \exp\left(\sqrt{\frac{1}{2}\log x \log\log x}\right).$$

Then, using the second part of (7), we deduce that for $\beta > 0$,

$$\Psi(x, L(x)^{\beta+o(1)}) \;=\; xL(x)^{-1/\beta+o(1)}. \tag{8}$$

From this one can easily deduce that

$$y_0(x) = L(x)^{1+o(1)}, \text{ and } J_0(x) = y_0^{2-\{1+o(1)\}/\log\log y_0} = L(x)^{2+o(1)}, \tag{9}$$

where y_0 and J_0 are as in the introduction (see (1)). From this we can deduce the following basic estimate, which we will use in later proofs.

LEMMA 2.1. *Fix constant $\beta > 0$. If $y = y_0^{\beta+o(1)}$, then*

$$\frac{\Psi(x,y)/y}{\Psi(x,y_0)/y_0} \;=\; y_0^{2-\beta-\beta^{-1}+o(1)}.$$

2.2. *Hildebrand-Tenenbaum saddle point method estimates.* For any $\alpha > 0$, one has

$$\Psi(x,y) \le \sum_{\substack{n\le x\\ P(n)\le y}} (x/n)^\alpha \le x^\alpha \xi(\alpha, y), \tag{10}$$

where

$$\xi(s,y) \;=\; \prod_{p\le y}\Big(1-\frac{1}{p^s}\Big)^{-1}.$$

Define $\alpha = \alpha(x,y)$ to be the solution to

$$\log x = \sum_{p\le y} \frac{\log p}{p^\alpha - 1}. \tag{11}$$

By [9, Th. 1 and (7.19)] we obtain in the range (6) with $u \to \infty$,

$$\Psi(x,y) \sim \frac{x^\alpha \xi(\alpha,y)}{\alpha\sqrt{2\pi \log x \log y}}. \tag{12}$$

Let $\xi = \xi(u)$ be the solution to $e^\xi = u\xi + 1$ so that

$$\xi(u) \;=\; \log(u\log u) + \frac{(1+o(1))\log\log u}{\log u} \text{as } u \to \infty. \tag{13}$$

Note also that $\xi'(u) \sim 1/u$. In the range (6), it turns out that

$$(1-\alpha(x,y))\log y = \xi(u) + O(1/u) \tag{14}$$

which implies that

$$y^{1-\alpha} = e^{\xi(u)}(1+O(1/u)) = u\xi(u)(1+O(1/u)). \tag{15}$$

So, for

$$y \;=\; L(x)^{\beta+o(1)} \;=\; y_0^{\beta+o(1)},$$

we have

$$y^{1-\alpha} \ \sim \ \beta^{-2}\log y \ \sim \ \beta^{-1}\log y_0. \tag{16}$$

By [9, Th. 3] and (14) above, we have

$$\Psi\left(\frac{x}{d},y\right) = \frac{\Psi(x,y)}{d^{\alpha(x,y)}}\left\{1+O\left(\frac{1}{u}+\frac{\log y}{y}\right)\right\} \text{ when } 1 \ \le \ d \ \le \ y \ \le \ \frac{x}{d}. \tag{17}$$

PROPOSITION 2.2. *There exists a constant $U>0$ such that throughout the range* (6) *with $x \ge y^U$, and for any $d \ge 1$, we have*

$$\Psi\left(\frac{x}{d},y\right) \le \frac{\Psi(x,y)}{d^{\alpha(x,y)}}\{1+o(1)\}$$

as $x\to\infty$, where α is the solution to (11). *In fact,*

$$\Psi\left(\frac{x}{d},y\right) \ < \ \frac{\Psi(x,y)}{d^{\alpha(x,y)}}$$

when $\log d \gg \log u\log y+\sqrt{u\log u\log y}$ and $u\ge U$; and

$$\Psi\left(\frac{x}{d},y\right) \ = \ \frac{\Psi(x,y)}{d^{\alpha(x,y)}}\left\{1+O\left(\frac{\log^2 u}{u}+\frac{\log(u+1)}{\log y}\right)\right\}$$

when $\log d \ll \log u\log y+\sqrt{u\log u\log y}$ or $u<U$.

Proof. If $d>x$, then the results are all trivial. Let $\nu = u/3\log u$. If $x \ge d \ge x/y^\nu$, then $d^{\alpha(x,y)}\Psi\left(\frac{x}{d},y\right) \le d^{\alpha(x,y)}(x/d) \le x/(x/y^\nu)^{1-\alpha(x,y)} = x/(y^{1-\alpha(x,y)})^{u-\nu} \asymp x/e^{(u-\nu)\xi(u)}$ by (15), which is $\ll \Psi(x,y)e^{-u/2}$ combining (13) with the first part of (7). This implies the Proposition for large d. Henceforth we may assume that $1\le d\le x/y^\nu$.

By (5), for $d=y^r$ with $0\le r\le u-\nu$, we have

$$\frac{\Psi\left(\frac{x}{d},y\right)d^\alpha}{\Psi(x,y)} \ = \ \frac{d^{-(1-\alpha)}\rho(u-r)}{\rho(u)}\ \left(1+O\left(\frac{\log(u+1)}{\log y}\right)\right).$$

The logarithm of the main term on the right side is

$$-(1-\alpha)r\log y+\log(\rho(u-r)/\rho(u)).$$

Using the fact that $u=(\log x)/(\log y)$, this can be rewritten as

$$r(\xi(u)-(1-\alpha)\log y)+\left(-\int_{u-r}^{u}\frac{\rho'(v)}{\rho(v)}dv-r\xi(u)\right).$$

The first term is $O(r/u)$ by (14). Corollary 8.3 of [15] gives that

$$-\rho'(v)/\rho(v)=\xi(v)(1+O(1/v)) \tag{18}$$

so that the second term equals

$$-\int_0^r(\xi(u)-\xi(u-t))dt+O\left((\log u)\log\frac{u}{u-r}\right).$$

Now, differentiating $e^\xi = u\xi + 1$, we obtain

$$\xi + u\xi' \;=\; \xi' e^\xi \;=\; \xi'(u\xi + 1)$$

so that

$$\xi' \;=\; \frac{1}{u - (u-1)\xi^{-1}} \;=\; \frac{1}{u(1 + O(1/\log u))} = \;\frac{1}{u}\left(1 + O\left(\frac{1}{\log u}\right)\right).$$

Therefore,

$$\begin{aligned}(19)\qquad \int_0^r (\xi(u) - \xi(u-t))dt &= \int_0^r (r-v)\xi'(u-v)dv \\ &= \left(1 + O\left(\frac{1}{\log u}\right)\right)\int_0^r \frac{(r-v)}{(u-v)}dv \\ &= \left(1 + O\left(\frac{1}{\log u}\right)\right)(r - (r-u)\log(1 - r/u)).\end{aligned}$$

Since

$$r - (r-u)\log(1-r/u) = \sum_{k=2}^{\infty} \frac{r^k}{k(k-1)u^{k-1}} = \frac{r^2}{2u}(1+A),\ 0 \le A \ll \frac{r}{u},$$

we obtain

$$\begin{aligned}\log\left(\frac{\Psi\left(\frac{x}{d}, y\right)d^\alpha}{\Psi(x,y)}\right) &= -\left(1 + O\left(\frac{1}{\log u}\right)\right)(r - (r-u)\log(1-r/u)) \\ &\quad +\; O\left((\log u)\log\left(\frac{u}{u-r}\right) + \frac{\log(u+1)}{\log y}\right) \\ &= \;-\frac{r^2}{2u}\left\{1 + A + O\left(\frac{1}{\log u} + \frac{u(\log u)\log\frac{u}{u-r}}{r^2}\right)\right\} \\ &\quad + O\left(\frac{\log(u+1)}{\log y}\right).\end{aligned}$$

The first claimed inequality follows. The last expression is negative provided that u is sufficiently large and $(\log u + \sqrt{u(\log u)/\log y})/r$ is sufficiently small. If u is bounded or if $r \ll \log u + \sqrt{u(\log u)/\log y}$, then this is $\ll \frac{\log^2 u}{u} + \frac{\log(u+1)}{\log y}$ in the range (6). □

We will require the following lemma, which is in one sense stronger, and in another sense weaker, than Lemma 2.1.

LEMMA 2.3. *We have*

$$\frac{\Psi(x,y)}{y} \;= o\left(\frac{\Psi(x,y_0)}{y_0(\log y_0)^{1+\varepsilon/4}}\right)$$

as $x \to \infty$, uniformly over y outside of the range

(20)

$$y_0 \exp(-(1+\varepsilon)\sqrt{\log y_0 \log\log y_0}) \;\le\; y \;\le\; y_0 \exp((1+\varepsilon)\sqrt{\log y_0 \log\log y_0});$$

and

$$\frac{\Psi(x,y)}{y} \ \le \ \frac{(2/e^2-\varepsilon)\Psi(x,y_0)}{y_0}$$

for all y outside of the range

$$y_0\exp(-(c_3+\varepsilon)\sqrt{\log y_0}) \ \le \ y \ \le \ y_0\exp((c_3+\varepsilon)\sqrt{\log y_0}). \tag{21}$$

Proof. Let $x = y_0^{u_0}$. Define $g(u) = g_x(u) = \log\rho(u) - u^{-1}\log x$. By (5) we have $\log(\Psi(x,y)/xy) = g(u) + O(1/u)$, provided $\log y \asymp \log L$. Select u_1 to maximize $g(u)$. Therefore $g(u_1) \ge g(u_0)$ by definition of u_1; and $g(u_0) \ge g(u_1) + O(1/u_0)$ by the definition of u_0 and the above estimate; therefore $g(u_0) = g(u_1) + O(1/u_0)$.

By (18), we have $g'(v) = \rho'(v)/\rho(v) + v^{-2}\log x = -\xi(v) + v^{-2}\log x + O(\log v/v)$; thus, for $t = O(u_1/\log u_1)$,

$$\begin{aligned} g'(u_1+t) &= g'(u_1+t) - g'(u_1) \\ &= \xi(u_1) - \xi(u_1+t) + \left(\frac{1}{(u_1+t)^2} - \frac{1}{u_1^2}\right)\log x + O\left(\frac{\log u_1}{u_1}\right) \\ &= O\left(\frac{t+\log u_1}{u_1}\right) - 2tu_1^{-3}\log x(1+O(t/u_1)) \\ &= -2t\frac{\xi(u_1)}{u_1} + O\left(\frac{t+\log u_1}{u_1}\right) \end{aligned}$$

since $0 = g'(u_1) = -\xi(u_1) + u_1^{-2}\log x + O(\log u_1/u_1)$. Therefore

$$g(u_1) - g(u_1+T) = -\int_0^T g'(u_1+t)dt = \frac{T^2}{u_1}(\xi(u_1)+O(1)) + O\left(\frac{T\log u_1}{u_1}\right) \tag{22}$$

for $T = O(u_1/\log u_1)$. We deduce that $u_0 = u_1 + O(1)$, as well as both

$$g(u) < g(u_0) - (1+\varepsilon/3)\log u_0 \text{ for } \ |u-u_0| > (1+\varepsilon/2)\sqrt{u_0}$$

and

$$g(u) < g(u_0) - \log(e^2/2+\varepsilon) \text{ for } \ |u-u_0| > (c_3+\varepsilon)\sqrt{u_0/\log u_0},$$

which are the desired results. □

Next we obtain a more accurate estimate for y_0 than (9).

LEMMA 2.4. *We have*

$$\log y_0 = \log L(x)\left(1 + \frac{\log_3 x - \log 2}{2\log_2 x} + O\left(\left(\frac{\log_3 x}{\log_2 x}\right)^2\right)\right)$$

and

$$\frac{u_0\xi(u_0)}{\log y_0} = 1 + O\left(\frac{1}{u_0}\right).$$

Proof. In the notation of the Lemma 2.3, we see by (22) that $|g(u_1+T)| = o(1/u_1)$ as $T \to \infty$ so that $u_0 = u_1+O(1)$. We saw that $u_1^2\xi(u_1)(1+O(1/u_1)) = \log x$, so the same equation is satisfied by u_0 (in place of u_1) and the estimate for $\log y_0 = (1/u_0)\log x$ follows from (13). Moreover

$$u_0\xi(u_0) = (\log y_0)(1 + O(1/u_0)). \qquad \square$$

COROLLARY 2.5. *If $d = p_1p_2\cdots p_k$, where each p_j is a prime in $(y_0, My_0]$, we have*

$$\frac{\psi(x/(p_1\cdots p_k),\, y_0)}{\psi(x, y_0)} \sim \frac{(\log y_0)^k}{p_1\cdots p_k} \tag{23}$$

uniformly in $1 \le k \le \log\log x$ and $\log M = o(\sqrt{(\log x)/(\log\log x)^3})$ as $x \to \infty$. Also,

$$\frac{\psi(x/(p_1\cdots p_k),\, y_0)}{\psi(x, y_0)} \le 2^k\frac{(\log y_0)^k}{p_1\cdots p_k} \tag{24}$$

uniformly for $\log M \le (\log y_0)/2\log\log y_0$, and all $k \ge 0$, for x sufficiently large.

Proof. Each $p_j \le y_0^2$. Hence, by the last part of Proposition 2.2, we have

$$\Psi\left(\frac{x}{p_1\cdots p_k}, y_0\right) \sim \frac{\Psi(x, y_0)}{(p_1\cdots p_k)^{\alpha(x,y_0)}}$$

for $k \le r \ll \log\log x$ (where $y_0^r = p_1\cdots p_k$). Now by (15) and the last part of Lemma 2.4, we know that $y_0^{1-\alpha(x,y_0)} = \log y_0(1 + O(1/u_0))$; hence

$$(p_1\cdots p_k)^{1-\alpha(x,y_0)} = M^{O(1-\alpha(x,y_0))}(\log y_0(1 + O(1/u_0)))^r \sim \log^r y_0.$$

Now $y_0^{r-k} \le M^k$ so that if $k\log M = o(\log y_0/\log\log y_0)$, then $\log^r y_0 \sim \log^k y_0$, as desired. (23) follows in the given range for M.

For the second part, note that by the first part of Proposition 2.2, we have

$$\Psi\left(\frac{x}{p_1\cdots p_k}, y_0\right) \le \{1 + o(1)\}\frac{\Psi(x, y_0)}{p_1\cdots p_k}(p_1\cdots p_k)^{1-\alpha(x,y_0)}.$$

By (16), $y_0^{1-\alpha} \sim \log y_0$, and so

$$1 \le \left(\frac{p_i}{y_0}\right)^{1-\alpha} \le M^{1-\alpha} = (\{1 + o(1)\}\log y_0)^{\log M/\log y_0} \le e^{1/2} + o(1).$$

Hence

$$\{1 + o(1)\}(p_1\cdots p_k)^{1-\alpha} \le \{1 + o(1)\}(\{e^{1/2} + o(1)\}\log y_0)^k \le (2\log y_0)^k$$

for y_0 sufficiently large and hence for x sufficiently large. $\square$

2.3. *Straightforward analytic estimates.* We complete this section by collecting together various straightforward analytic estimates that will be needed later.

Fix $0 < a < b$. By the prime number theorem, we have

$$\sum_{ay<q\le by} \frac{\log y}{q} \sim \log\left(\frac{b}{a}\right), \tag{25}$$

where the sum is over primes q, and also that

$$\sum_{ay<q\le by} \frac{\log y}{q} \le 2\log\left(\frac{b}{a}\right), \tag{26}$$

for all $1 \le a \le b/2$, once y is sufficiently large. To see this note that, since $\sum_{q\le Q}(\log q)/q = \log Q + C + o(1)$, for some constant C, the sum is

$$\le \sum_{ay<q\le by} \frac{\log q}{q} = \log\left(\frac{b}{a}\right) + o_{y\to\infty}(1),$$

and the result follows.

LEMMA 2.6. *Let*

$$g(\beta, C) \;:=\; \beta^{-2}\int_0^{C/\beta^2} \log\left(\frac{e^z+e^{-z}}{2}\right)\frac{dz}{z^2} + 1 - \log(C). \tag{27}$$

The function $g(1, C)$ is decreasing for $C > 0$, with

$$\lim_{C\to\infty} g(1, C) = \gamma + \log(4/\pi).$$

Proof. Since

$$\frac{dg(1,C)}{dC} = \frac{\log(\frac{1}{2}(e^C+e^{-C}))}{C^2} \;-\; \frac{1}{C} \;< 0,$$

for all $C > 0$, we minimize by letting $C \to \infty$. Integrating by parts, we have that

$$\lim_{C\to\infty} g(1, C) = \int_0^1 \frac{e^z - e^{-z}}{e^z+e^{-z}}\frac{dz}{z} - 2\int_1^\infty \frac{e^{-z}}{e^z+e^{-z}}\frac{dz}{z}.$$

Now 6.1.50 of [1] states that

$$\log\Gamma(s) = \int_0^\infty \left((s-1)e^{-t} - \frac{e^{-t}-e^{-st}}{1-e^{-t}}\right)\frac{dt}{t};$$

and the third line of 6.3.22 of [1] readily implies that

$$\gamma = \int_0^1 (1-e^{-t})\frac{dt}{t} - \int_1^\infty e^{-t}\frac{dt}{t}. \tag{28}$$

Since $\Gamma(1/2) = \pi^{1/2}$, and taking $s = 1/2$ and $t = 4z$, our result follows. □

3. The lower bound for T in Theorems 1.2 and 1.3

3.1. *Proof strategy.* To establish that

$$\text{Prob}\Big(T \ > \ (\pi/4)(e^{-\gamma}-\varepsilon)J_0(x)\Big) \ = \ 1-o(1),$$

we show that the expected number of nontrivial subsets S of $\{1,\dots,J\}$ for which $\prod_{i\in S} a_i$ is a square is $o(1)$ for $J(x)=(\pi/4)(e^{-\gamma}-\varepsilon)J_0(x)$.

3.2. *Structure of a square product.* We begin with the following proposition.

PROPOSITION 3.1. *Select integers $a_1,\dots,a_J$ at random from $[1,x]$. The probability that there exists a subsequence I of the a_i with*

$$2 \ \le \ |I| \ \le \ \frac{\log x}{2\log\log x} \quad \textit{for which} \ \prod_{a\in I} a \ \textit{is a square}$$

is $O(J^2\log x/x)$ provided $J<x^{o(1)}$.

Proof. Suppose that $b_1,\dots,b_k$ were chosen at random from $[1,x]$. The probability that $b_1b_2\dots b_k$ is a square equals

$$x^{-k}|\{b_1,\dots,b_k \ \le \ x \ : \ b_1b_2\dots b_k \text{ is a square}\}|.$$

Now write each b_i uniquely as

$$b_i \ = \ c_iu_i^2, \text{ where } c_i \text{ is squarefree.}$$

Assuming that $b_1\cdots b_k$ is a square, which implies $c_1\cdots c_k$ is a square, define the doubly indexed sequence $c_{i,j}$, where $i,j=1,\dots,k$ and $i\ne j$, to be any satisfying the relations

$$c_{i,j} \ = \ c_{j,i}, \quad \text{with} \ \ c_i \ = \ \prod_{j\ne i} c_{i,j} \ \ \text{for each } i. \tag{29}$$

The fact that such $c_{i,j}$ exist can be seen as follows. For each prime p dividing $c_1\cdots c_k$, we will need to decide which $c_{i,j}$ that p divides; and, to do this, suppose that p divides $c_{i_1},\dots,c_{i_{2t}}$. (The reason it is $2t$ is that all the c_i are square-free and have product a square.) Then, the following $c_{i,j}$ are to be divisible by p, and no others:

$$c_{i_1,i_2}, c_{i_2,i_1}, c_{i_3,i_4}, c_{i_4,i_3}, \dots, c_{i_{2t-1},i_{2t}}, c_{i_{2t},i_{2t-1}}.$$

Each $c_{i,j}$ is then the product of the primes dividing $c_1\cdots c_k$ that divide it; and if this process leaves some $c_{i,j}$ not divisible by any prime $p|c_1\cdots c_k$, then we set $c_{i,j}=1$.

Given $c_1,\dots,c_k$, the number of sequences $b_1,\dots,b_k$ satisfying $b_i=c_iu_i^2$ is the number of possibilities for the numbers u_i, which is $\le (x/c_i)^{1/2}$; and so,

the probability that $b_1 \cdots b_k$ is a square is

$$\leq \frac{1}{x^k} \sum_{\substack{c_{i,j} \leq x \\ \text{for } 1 \leq i < j \leq k}} \prod_{i=1}^{k} \left(\frac{x}{\prod_{j \neq i} c_{i,j}} \right)^{1/2} \tag{30}$$

$$\leq \frac{1}{x^{k/2}} \sum_{1 \leq i < j \leq k} \left(\prod_{c_{i,j} \leq x} \frac{1}{c_{i,j}} \right) \leq \frac{1}{x^{k/2}} \, (1 + \log x)^{k(k-1)/2}$$

since each $c_{i,j}$ appears twice in the above product. Therefore the probability that there exists $I \subset \{1, 2, \ldots, J\}$ for which $\prod_{i \in I} a_i$ is a square, with $|I| = k$, is

$$\leq \binom{J}{k} \frac{1}{x^{k/2}} \, (1 + \log x)^{k(k-1)/2} \leq \left(\frac{J^2 (1 + \log x)^{k-1}}{x} \right)^{k/2},$$

which gives $O(J^2 \log x/x)$ for $k = 2$ and is $\leq 1/x$ for $3 \leq k \leq \log x / 2 \log\log x$. □

3.3. *The main argument.* In this subsection, we prove that

$$\text{Prob}\Big(T \ > \ (\pi/4)(e^{-\gamma} - \varepsilon) J_0(x) \Big) \ = \ 1 - o(1).$$

As a consequence of the upper bound proved in [5], we may assume that $T < (3/4)J_0(x)$ holds with probability $1 - o(1)$. Furthermore, following Proposition 3.1, we need only focus on subsequences I of $a_1, \ldots, a_J$ (where $J = T < J_0(x)$) of length exceeding $\log x / 2 \log\log x$ that have product equal to a square.

Throughout we shall write $a_i = b_i d_i$ where $P(b_i) \leq y$ and where either $d_i = 1$ or $p(d_i) > y$ for $1 \leq i \leq k$. In this subsection we shall select $y = J_0(x)^{O(1)}$. Recall here that $p(n)$ denotes the smallest and $P(n)$ the largest prime divisor of n. If $a_1, \ldots, a_k$ are chosen at random from $[1, x]$, then

$$\text{Prob}(a_1 \ldots a_k \text{ is a square}) \leq \text{Prob}(d_1 \ldots d_k \text{ is a square}) \tag{31}$$

$$= \sum_{\substack{d_1, \ldots, d_k \geq 1 \\ d_1 \ldots d_k \text{ is a square} \\ d_i = 1 \text{ or } p(d_i) > y}} \prod_{i=1}^{k} \frac{\Psi(x/d_i, y)}{x}$$

$$\leq \left(\{1 + o(1)\} \frac{\Psi(x,y)}{x} \right)^k \sum_{n=1 \text{ or } p(n) > y} \frac{\tau_k(n^2)}{n^{2\alpha}}$$

by Proposition 2.2, where $\tau_k(m)$ denotes the number of different ways of writing m as the product of k positive integers.

Out of $J = \eta J_0$ integers, the number of k-tuples is $\binom{J}{k} \le (eJ/k)^k$; and so the expected number of k-tuples whose product is a square is

$$\le \left((e+o(1))\frac{\eta y}{k\log y_0}\frac{\Psi(x,y)/y}{\Psi(x,y_0)/y_0}\right)^k \prod_{p>y}\left(1+\frac{\tau_k(p^2)}{p^{2\alpha}}+\frac{\tau_k(p^4)}{p^{4\alpha}}+\dots\right). \tag{32}$$

We now consider k in two different ranges, and in both ranges we will select different values for y so as to give good upper bounds for (32).

- First, if

$$\frac{\log x}{2\log\log x} \;<\; k \;\le\; y_0^{1/4},$$

then let $y = y_0^{1/3}$ so that $k = o(y_0^{\alpha})$. Therefore the Euler product in (32) is

$$\le \exp\left(O\left(\sum_{p>y}\frac{k^2}{p^{2\alpha}}\right)\right) \le \exp\left(O\left(\frac{k^2 y^{2(1-\alpha)}}{y\log y}\right)\right) = e^{o(k)}.$$

Now $\Psi(x, y_0^{\gamma}) = x/y_0^{1/\gamma+o(1)}$ by (8), and therefore the quantity in (32) is

$$\le \left(\frac{1/y_0^{3+o(1)}}{k/y_0^{2+o(1)}}\right)^k \le y_0^{-k+o(k)}, \tag{33}$$

which is $< 1/x^2$ in this first range for k.

- Next, we consider the range

$$y_0^{1/4} \;\le\; k \;=\; y_0^{\beta} \;\le\; J \;\le\; J_0.$$

In this case we will choose y so that $[k/C] = \pi(y)$; then we will optimize the C later. For this choice of y, a simple calculation reveals that

$$\begin{aligned}\frac{\tau_k(p^2)}{p^{2\alpha}}+\frac{\tau_k(p^4)}{p^{4\alpha}}+\cdots \;&\sim\; \frac{(k/p^{\alpha})^2}{2!}+\frac{(k/p^{\alpha})^4}{4!}+\cdots \\ &=\; \frac{e^{k/p^{\alpha}}+e^{-k/p^{\alpha}}}{2}-1.\end{aligned}$$

In order to evaluate (32), we need to take the product of this over the primes $p > y$. The logarithm of this product equals

$$\sum_{\substack{p>y\\ p \text{ prime}}} \log\left(\frac{e^{k/p^{\alpha}}+e^{-k/p^{\alpha}}}{2}\right) \;\sim\; \int_y^{\infty}\frac{1}{\log t}\log\left(\frac{e^{k/t^{\alpha}}+e^{-k/t^{\alpha}}}{2}\right)dt$$

by the prime number theorem. Letting $z = k/t^{\alpha}$, from (16) this last integral is

$$\sim \int_0^{C/\beta^2}\frac{(k/z)^{1/\alpha}}{z\log(k/z)}\log\left(\frac{e^z+e^{-z}}{2}\right)dz.$$

Now, $k^{1/\alpha-1} \sim \beta^{-2}\log y$ by (16) so that

$$\frac{(k/z)^{1/\alpha}}{\log(k/z)} \ \sim \ (k/z)\beta^{-2}$$

as $z = O(1)$. It follows that the quantity in (32) is bounded from above by

$$\left((1+o(1))e^{g(\beta,C)}\beta\eta\frac{\Psi(x,y)/y}{\Psi(x,y_0)/y_0}\right)^k, \tag{34}$$

where $g(\beta, C)$ is defined in (27).

Now, for any fixed C we have, as a consequence of Lemma 2.1, that (34) is $o(1/x^2)$ unless $\beta = 1 + o(1)$; and so, we really only need to consider $k = y_0^{1+o(1)}$, as the total expected number of k-tuples for other values of k add only $o(1/x^{2+o(1)})$. If $C = C(\varepsilon)$ is sufficiently large, then $e^{g(1,C)} < 4e^\gamma/\pi + \varepsilon$ by Lemma 2.6 and, since y_0 maximizes $\Psi(x,y)/y$ for $y = y_0$, we deduce that (32) is at most

$$\le ((1+\varepsilon)4\eta e^\gamma/\pi)^k.$$

Therefore, if $\eta < (1-\varepsilon)e^{-\gamma}\pi/4$, then this is less than $1/x^2$. This finishes the proof of the lower bound in Theorem 1.2. □

3.4. *Proof of Theorem* 1.3, *part* (a). This last proof yields further useful information: If either $J < (\pi/4)(e^{-\gamma}-\varepsilon)J_0(x)$, or if $k < y_0^{1-o(1)}$ or $k > y_0^{1+o(1)}$, then the expected number of square products with $k > 1$ is $O(J_0(x)^2\log x/x)$, whereas the expected number of squares in our sequence is $\sim J/\sqrt{x}$. This justifies Theorem 1.3(a).

3.5. *Proof of Theorem* 1.3, *part* (b). The proof in Section 3.3 yielded that if we have a square product then, with probability $1 + o(1)$, we have $|I| = k = y_0^{1+o(1)}$. We now assume that $k = y_0^{1+o(1)}$ with

$$k \ \notin \ [y_0\exp(-(c_3+\varepsilon)\sqrt{\log y_0}),\ y_0\exp((c_3+\varepsilon)\sqrt{\log y_0})]. \tag{35}$$

From the discussion following (34) above, we know, by taking C large, that the expected number of such k-tuples is at most

$$\left((4e^\gamma/\pi+\varepsilon)\frac{\eta\Psi(x,y)/y}{\Psi(x,y_0)/y_0}\right)^k.$$

By Lemma 2.3, this is at most

$$\Big((4e^\gamma/\pi+\varepsilon)(2/e^2+o(1))\eta\Big)^k \ < \ 1/2^k$$

for sufficiently small $\varepsilon > 0$, using the fact that $\eta < 3/4$. Therefore the expected number of k- tuples with product a square is $o(1)$ for all k satisfying (35), so that Theorem 1.3(b) follows. □

3.6. *Proof of Theorem* 1.3, *part* (c). In the previous subsection we proved that

$$|I| \ \le \ y_1 \ := \ y_0 \exp((1+\varepsilon)\sqrt{\log y_0 \log\log y_0}),$$

with probability $1-o(1)$. In this section we prove, among other results, part (c) of Theorem 1.3.

PROPOSITION 3.2. *Write each* $a_i = b_i d_i$ *where* $P(b_i) \le y = y_1 < p(d_i)$, *and suppose that* $d_{i_1} \dots d_{i_l}$ *is a subproduct which equals a square* n^2, *but such that no subproduct of this is a square. Then, with probability* $1-o(1)$, *for all such products, we have* $l = o(\log y_0)$ *and* n *is a squarefree integer composed of precisely* $l-1$ *prime factors, each* $< y^2$, *where* $n \le y^{2l}$.

Proof. For ease of notation we will relabel, replacing $d_{i_1} \dots d_{i_l}$ by $d_1 \dots d_l$. Note that with the choice of $y = y_1$, we have $y/l \log y \to \infty$ and $y = y_0^{1+o(1)}$, so we know that $y^\alpha \sim y/\log y$ by (16).

We now show that n has at least $l-1$ (not necessarily distinct) prime factors so that $n^2 = d_1 \dots d_l > y^{2(l-1)}$. Let d_j' be the product of the primes which divide d_j to an odd power so that each d_j' is squarefree. Then $n^2 \ge d_1' \dots d_l' = N^2$ and no proper subproduct can be a square. We now create a graph G on l vertices $v_1, \dots, v_l$, representing $d_1', \dots, d_l'$, respectively. For each prime p dividing N, suppose that p^k is the highest power of p dividing N. Since each d_j' is squarefree, and p^{2k} is the highest power of p dividing their product, it follows that $J_p := \{j : \ p | d_j'\}$ has exactly $2k$ elements. We now create a *perfect matching* of k edges, colored p, in the subgraph $\{v_j : j \in J_p\}$ of $2k$ vertices, in G; that is, we draw k edges colored p in G, with each v_j with $j \in J_p$ on the end of one such edge (and this matching can be done arbitrarily within these constraints). Hence we derive each v_j from the corresponding d_j' and, vice-versa, d_j' is the product of those primes p for which there is an edge of color p adjacent to v_j. Now, the product over the (d_j' corresponding to the) vertices in each *connected component* of such a graph is a square since each edge corresponds to the square of a prime. By the minimality of the product $\prod_j d_j'$, there can therefore only be one component, meaning that the graph is connected. Hence the graph has at least $l-1$ edges, implying that there are at least $l-1$ (not necessarily distinct) primes dividing N and hence that $n \ge N > y^{l-1}$.

We now modify the argument from the start of Section 3.3 (with k replaced by l) to restrict our attention to cases in which $d_1 \dots d_l \ge y^{2l}\phi(x)^2$, where $\phi(x) = y^{O(1)}$. To obtain an upper bound we may multiply through the summand, in (31), by $(n/y^l\phi(x))^{2\theta}$, where we have chosen $\theta > 0$ so that $y^{2\theta} = (2y\log l)/(l(\log y)^2)$. Then we must multiply the right side of (32) through by $1/(y^{2\theta})^l\phi(x)^{2\theta}$ and change the terms in the Euler product to $(1 + \tau_l(p^2)/p^{2(\alpha-\theta)} + \tau_l(p^4)/p^{4(\alpha-\theta)} + \cdots)$.

First we bound the Euler product using the prime number theorem. Recall that the function $\tau_l(n)$ counts the number of sequences of positive integers $d_1, \dots, d_l$ such that $d_1 \cdots d_l = n$. In the case $n = p^{2k}$, this amounts to computing the number of ordered partitions of $2k$ into l parts that are ≥ 0; so,

$$\tau_l(p^{2k}) = \binom{2k+l-1}{2k} \leq \begin{cases} l(l+1)/2, & \text{if } k=1, \\ \frac{(2k+l-1)^{2k}}{(2k)!,} & \text{if } k \geq 2. \end{cases}$$

For $p = y_0^{1+o(1)} = L(x)^{1+o(1)}$, using (16) with $\beta = 1$, we have that

$$\frac{1}{p^{2\alpha}} \sim \frac{\log^2 p}{p^2},$$

making the summation of terms involving p in the Euler product become

$$\{1+o(1)\}\frac{l(l+1)}{2} \cdot \frac{\log^2 p}{p^2} \cdot p^{2\theta}.$$

Via the prime number theorem, the logarithm of the Euler product is therefore

$$\sim \frac{l(l+1)}{2} \sum_{y<p<y^4} \frac{\log^2 p}{p^{2-2\theta}} \sim \frac{l(l+1)}{2} \int_y^{y^4} \frac{\log t}{t^{2-2\theta}} dt.$$

(Here the primes p, with $y < p < y^{4+o(1)}$, being the only relevant ones follows from comments made above the statement of Theorem 1.3.) Now $\theta < 1/2$ by definition, so the above calculation becomes

$$\sim \frac{l(l+1)}{2} \cdot \frac{\log y}{(1-2\theta)y^{1-2\theta}} = \frac{l(l+1)}{2} \cdot \frac{\log^2 y}{y^{1-2\theta}(1-2\theta)\log y}.$$

Now $y^{1-2\theta} = l \log^2 y / 2 \log l$, so the above is

$$= \frac{(l+1)\log l}{\log\left(l \log^2 y / 2 \log l\right)} \leq l+1$$

since $l \leq y$. So putting (32) to use as explained above, the expected number of such l-tuples is

$$\leq \frac{1}{\phi(x)^{2\theta}} \left((e+o(1)) \, \frac{\eta y}{l y^{2\theta} \log y_0} \, \frac{\Psi(x,y)/y}{\Psi(x,y_0)/y_0} \right)^l e^{l+1} \tag{36}$$

$$= \frac{1}{\phi(x)^{2\theta}} \left((e+o(1)) \, \frac{\eta(\log^2 y)}{(2\log l)(\log y_0)} \, \frac{\Psi(x,y)/y}{\Psi(x,y_0)/y_0} \right)^l e^{l+1} \tag{37}$$

$$\ll \frac{1}{\phi(x)^{2\theta}(\log y_0)^{\varepsilon l/5}} \tag{38}$$

as $\eta \leq 1$ and by Lemma 2.3 for $y = y_1$.

Now we are ready to establish the conclusions of the proposition. Take $\phi(x) = 1/y$ in the above, and as $2\theta < 1$ by definition, (38) becomes $\ll y/(\log y)^{\varepsilon l/5}$. This is $o(1)$ provided $l \geq 6 \log y/(\varepsilon \log\log y)$, hence we expect

$o(1)$ products with $l \gg \log y_0$, yielding $l = o(\log y_0)$ with probability $1 - o(1)$. In this case, $2\theta \sim 1$.

Regarding the structure of the factorization of n: Taking $\phi(x) = 1$, we expect $o(1)$ products with $d_1 \dots d_l \geq y^{2l}$; hence $d_1 \dots d_l = n^2 < y^{2l}$ with probability $1 - o(1)$. Since each prime divisor is $> y$, evidently n has $< l$ prime factors, and so exactly $l - 1$. Also, if p is the largest, then $y^{l-2}p < y^l$; that is, $p < y^2$.

Finally, we are left with showing that n is squarefree. To obtain an upper bound on the expected number of square products n^2 for which n is divisible by the square of a prime $> y$, we proceed much as above with $\phi(x) = 1/y$, but now the Euler product has an additional factor

$$\sum_{p>y} \left(\frac{\tau_l(p^4)}{p^{4\alpha-4\theta}} + \frac{\tau_l(p^6)}{p^{6\alpha-6\theta}} + \cdots \right) \ll \frac{l^4 (\log y)^3}{y^{3-4\theta}} = \frac{(2l \log l)^2}{y \log y}.$$

From (38) we thus deduce that we expect $o(1)$ such square products. □

4. **Hypergraphs**

The main result of this section is to prove the upper bound in Theorem 1.2. A roadmap for the proof is as follows.

Recall that the numbers $a_1, a_2, \dots$, chosen uniformly at random from $\{1, 2, \dots, x\}$, are encoded as row vectors over $\mathcal{F}_2$. Subsets whose product is a square are determined by combinatorial relations among these row vectors. Schroeppel's method, and its variants ignore columns corresponding to primes less than y_0. This makes the relations easier to satisfy, but we pay for it by requiring $\pi(y_0)$ many relations. To make the search more tractable, we restrict our attention to the more obvious ways of finding linear relations. Schroeppel's original method considers only the most obvious: after removing columns less than y_0, we must be left with all zeros. The *one large prime* variation considers also the next most obvious: when we have two identical rows containing a single 1.

The upper bound in Theorem 1.2 is proved via the *many large primes variation*. Tractability of the analysis rests on the fact that the combinatorial structure converges as $x \to \infty$ to a random object built from a Poisson point process. In order for the convergence to be uniform, we must restrict the columns: specifically, fixing $M > 1$, we must not use any a_i with a prime factor greater than My_0. We must also restrict the combinatorial complexity of the search for linear relations as follows. Calling two rows "neighbors" if they share a nonzero column entry (whose index is now forced to be between y_0 and My_0), any linear relation must take place within a ball of some fixed radius m in the neighbor graph on rows. We may then prove that the combinatorial structure converges in an appropriate sense to a tree-like random hypergraph defined on

a Poisson point process. The number of samples needed to accumulate $\pi(y_0)$ linear relations in the limiting model is computable explicitly in terms of some functions $\gamma_{m,M}$. For fixed m and M, these are ugly, but as $m, M \to \infty$, this number decreases to $e^{-\gamma} J_0$.

An outline of this section is as follows. Section 4.1 defines some functions that include the family $\{\gamma_{m,M}\}$. A result (Theorem 4.1) is then formulated in terms of these functions that implies the upper bound in Theorem 1.2. The subsection ends with the definition of some combinatorial structures such as tree-like hypergraphs that will be used in the search for linear relations. Section 4.2 formally defines the probability model and the random objects (hypergraphs with distinguished vertices) that will witness linear relations. The number of rows neighboring any given row is shown to have finite first and second moments (Proposition 4.3), which is then parlayed into an upper bound on the mean of size of the m-ball in the neighbor graph on rows. Section 4.3 constructs the limit object, an informal description of which appears at the beginning of that subsection. Section 4.4 proves convergence of the random hypergraphs in Section 4.2 to the limit object of Section 4.3. Although it takes several pages, it consists merely of repeated applications of Proposition 4.3. Section 4.5 evaluates the probability $\theta_m^{M,\eta}(\rho)$, which is the probability in the limit model that if a row containing a single 1 in column ρy_0 arises at time ηJ_0, it will form a new linear relation. The key result here (Lemma 4.18) is that this is 1 when m and M are sufficiently large and $\eta > e^{-\gamma}$. Finally, Section 4.6 finishes the proof of the main theorems.

4.1. *Preliminary results.* To begin in earnest, we define the following functions, which will arise in the branching processes with finite values of m and M:

$$A_M(z) := \int_{1/M}^{1} \frac{1 - e^{-zt}}{t}\, dt.$$

Clearly, as $M \to \infty$, we have the limit

$$A_M(z) \uparrow A(z) := \int_{0}^{1} \frac{1 - e^{-zt}}{t}\, dt.$$

Recursively, define functions $\gamma_{m,M}$ for $m = 0, 1, 2, \ldots$ by

$$\begin{aligned} \gamma_{0,M}(u) &:= u, \\ \gamma_{m+1,M}(u) &:= u \exp\left[A_M(\gamma_{m,M}(u))\right]. \end{aligned} \tag{39}$$

Note that $\gamma_{m,M}(u)$ is increasing in all three arguments. From this it follows that $\gamma_{m,M}(u)$ increases to $\gamma_M(u)$ as $m \to \infty$, a fixed point of the map $z \mapsto u \exp(A_M(z))$, so that

$$\gamma_M(u) := u \exp\left[A_M(\gamma_M(u))\right]. \tag{40}$$

We now establish that $\gamma_M(u) < \infty$ except perhaps when $M = \infty$. We have $0 \le A_M(z) \le \log M$ for all z, so that $u < \gamma_M(u) \le Mu$ for all u; in particular $\gamma_M(u) < \infty$ if $M < \infty$. As $M \to \infty$, the fixed point $\gamma_M(u)$ increases to the fixed point $\gamma(u)$ of the map $z \mapsto ue^{A(z)}$, or to ∞ if there is no such fixed point, in which case we write $\gamma(u) = \infty$. In Lemma 4.18 we show that this map has a fixed point if and only if $u \le e^{-\gamma}$. Otherwise $\gamma(u) = \infty$ for $u > e^{-\gamma}$ so that

$$\int_0^\eta \frac{\gamma(u)}{u}\, du = \infty > 1 \tag{41}$$

for any $\eta > e^{-\gamma}$.
Our main result in this section is the following.

THEOREM 4.1. *If η, m, M are such that*

$$\int_0^\eta \frac{\gamma_{m+1,M}(u)}{u}\, du > 1,$$

then with probability approaching 1, *as $x \to \infty$, among ηJ_0 uniform random samples from $\{1, \dots, x\}$, the subset of numbers that are (My_0)-smooth will contain a square subproduct. Furthermore, this will be witnessed in diameter at most m, in a sense to be made precise in Definitions* 4.7 *and* 4.9 *below.*

Together with (41), this establishes the upper bound in Theorem 1.2. Our conjecture that the upper bound is sharp is supported by the fact that $\lim_{t \uparrow \eta_*} \int_0^t \frac{\gamma(u)}{u}\, du = 1$, where $\eta_* := e^{-\gamma}$.

Hypergraphs. A *hypergraph* on a vertex set V is simply a collection $\mathcal{H}$ of finite subsets of V of cardinality at least 2. Each $S \in \mathcal{H}$ is called a *hyperedge* of $\mathcal{H}$; the *cardinality* of a hyperedge S is its cardinality as a set. Define the *support* of a hypergraph $\mathcal{H}$, denoted by $\operatorname{supp}(\mathcal{H}) := \bigcup_{S \in \mathcal{H}} S$, to be the union of all of its hyperedges. By a hypergraph $\mathcal{H}$ with vertex set V, we mean that $\operatorname{supp}(\mathcal{H}) \subseteq V$. (Note: in the literature, often this language would imply $\operatorname{supp}(\mathcal{H}) = V$.) We will typically use script letters for hypergraphs: $\mathcal{G}, \mathcal{H}$, and so forth. A rooted hypergraph is simply a hypergraph together with a choice of a distinguished element in its support. Thus, the hypergraphs on V rooted at p are in one-to-one correspondence with hypergraphs on V containing p in their support.

Definition 4.2 (tree-like hypergraphs). A finite hypergraph $\mathcal{G}$ rooted at p is *tree-like* if $\operatorname{supp}(\mathcal{G})$ may be given the structure of a tree T, rooted at p, in such a way that the following decomposition holds. Let I denote the set of vertices that are not leaves of T. We require that for each $q \in I$, the set of children of q may be partitioned into sets $V_{q,1}, \dots, V_{q,n(q)}$ so that each hyperedge of $\mathcal{G}$ is equal to $V_{q,j} \cup \{q\}$ for a unique pair (q, j) with $q \in I$ and $j \le n(q)$.

A moment's thought shows that if $\mathcal{G}$ is a tree-like hypergraph rooted at p, then the tree structure on $\operatorname{supp}(\mathcal{G})$ satisfying the definition is unique (when p is specified as the root). Denote this tree by $\mathbf{T}_p(\mathcal{G})$.

Sometimes it will be desirable to allow singleton hyperedges (hyperedges consisting of a single vertex, p). Rather than change the definitions, we introduce the notion of a *marked hypergraph*. This is just a pair $(\mathcal{G}, U)$, where $\mathcal{G}$ is a finite hypergraph and U is any subset of $\operatorname{supp}(\mathcal{G})$. We think of U as telling us (by marking) which singleton edges $\{p\}$ have been added to $\mathcal{G}$. Hypergraphs $\mathcal{G}$ and $\mathcal{G}'$ are defined to be isomorphic if there is a bijection $\phi : \operatorname{supp}(\mathcal{G}) \to \operatorname{supp}(\mathcal{G}')$ inducing a bijection at the level of hyperedges. Marked hypergraphs $(\mathcal{G}, U)$ and $(\mathcal{G}', U')$ are isomorphic if ϕ can be chosen so that also $\phi(U) = U'$.

In what follows, we will require a notion of weak convergence of probability measures on hypergraphs and marked hypergraphs, which in turn requires a metric on the space of marked hypergraphs on the vertex set $\mathbb{R}$ rooted at p. (We will re-normalize, replacing prime p by the real number $\rho = \rho_p := p/y$, which will thus lie in the fixed interval $(1, M]$.) It will turn out that all but a vanishing fraction of our hypergraphs are tree-like, so we need only to define the metric on tree-like hypergraphs (e.g., by convention we take the distance between hypergraphs to be $+\infty$ if either one is not tree-like). If $\mathcal{G}$ and $\mathcal{H}$ are two tree-like hypergraphs, define the distance to be $+\infty$ if the two hypergraphs are not isomorphic, and otherwise define the distance to be the least $\varepsilon > 0$ such that there is a bijection $\phi : \operatorname{supp}(\mathcal{G}) \to \operatorname{supp}(\mathcal{H})$ inducing an isomorphism on the hypergraphs, and satisfying $|\phi(\rho) - \rho| \le \varepsilon$ for all $\rho \in \operatorname{supp}(\mathcal{G})$. (Here we are dealing with re-normalized values of p; that is, $\rho_p = p/y$, which are bounded.) In other words, the topology is discrete on the graph structure along with the product topology on the names of the vertices. Formally,

$$d(\mathcal{G}, \mathcal{H}) := \min_{\phi} \left\{ \max_{\rho \in \operatorname{supp}(\mathcal{G})} |\phi(\rho) - \rho| \ : \ \phi \text{ is an isomorphism from } \operatorname{supp}(\mathcal{G}) \text{ to } \operatorname{supp}(\mathcal{H}) \right\}.$$

Define the distance between marked hypergraphs similarly, with ϕ now restricted to isomorphisms of the marked hypergraphs. Let μ and μ' be two probability measures on the space of hypergraphs on the vertex set $\mathbb{R}$. Say that a random pair $(\mathcal{G}, \mathcal{G}')$ of hypergraphs is a coupling of μ and μ' when $\mathcal{G}$ has law μ and $\mathcal{G}'$ has law μ'. Define the distance $d(\mu, \mu')$ between the probability measures μ and μ' to be the infimum of values $\varepsilon > 0$ such that there is a coupling $(\mathcal{G}, \mathcal{G}')$ of μ and μ' for which the probability of $d(\mathcal{G}, \mathcal{G}') > \varepsilon$ is at most ε. This is a standard metrization of the weak topology, that is, $d(\mu_n, \mu) \to 0$ if and only if $\int f \, d\mu_n \to \int f \, d\mu$ for all bounded and weakly continuous functions f.

Remark. Let $\pi_k(\mu)$ denote the restriction of μ to the subspace $\mathbf{H}_k$ of hypergraphs all of whose hyperedges have cardinality at most k. Weak convergence of μ_n to μ is equivalent to $\pi_k(\mu_n) \to \pi_k(\mu)$ for all k, together with *tightness*: $\mu_n(\mathbf{H}_k^c) \leq g(k)$ for all n, where $\lim_{k\to\infty} g(k) = 0$.

4.2. *The random hypergraph $\mathcal{G}$ of (My_0)-smooth numbers.* Before we get started, here are a few words on notation. As before, we are selecting random positive integers in $\{1, \ldots, x\}$ with $y = y_0(x)$ and $J_0(x)$ as in Section 1. Also, as before, we will choose an integer $J := \lfloor \eta J_0 \rfloor$ for some $\eta > 0$. We will choose a real $M > 1$ and keep track of large prime factors in the interval (y, My). By the term *large prime*, we will mean a prime in the interval (y, My). We will specify an integer $m \geq 1$ that is interpreted as the maximum chain length our algorithm will exploit when counting pseudosmooths, where a chain is a sequence $a_1, a_2, \ldots, a_r$, $r \leq m$, such that each consecutive pair a_i, a_{i+1} share a large prime factor $p_i \in (y, My)$. The first mission of this subsection is to define a random hypergraph that will depend on M, J, x, m and a large prime $p \in (y, My)$. The full notation for this will be $\mathcal{G}_{m,p}^{M,J,x}$. However, in most of the results and constructions that follow, M and J are fixed and x is a size parameter fixed during each construction, while m and p are dynamic. (The constructions are recursive in m and p and the proofs inductive.) Because of this, we often reduce clutter in the notation by writing simply $\mathcal{G}_{m,p}$ with the other three parameters understood. In many of our lemmas, the following phrase arises: "$f = o(1)$ as $x \to \infty$, uniformly as M and η vary over bounded intervals and $y < p < My$." To be precise about this once and for all, it means that there is a function g, going to zero as x goes to infinity, such that $f(M, J, x, m, p) < g(M_0, \eta, x, m)$ for all $M \leq M_0, J \leq \eta J_0$ and $y < p < My$ as $x \to \infty$. This holds for any fixed m, M_0, η. Several times in Section 4.4 below we prove weak convergence results. (Note: such convergence results needing to be uniform, in the manner just described, was the reason for metrizing the weak topology.)

Now we move on to the constructions. Fix an integer $x > 0$, and let $(\Omega_x, \mathcal{F}_x, \mathbb{P}_x)$ be a probability space on which is defined a sequence $\{X_1, X_2, \ldots\}$ of IID random variables whose common distribution is uniform on the set $\{1, 2, \ldots, x\}$. Let $y = y_0(x)$ and $J_0(x) = x\pi(y)/\psi(x, y)$ be as in Section 1. For each real $M > 1$ and each integer $J > 0$, we will define a random hypergraph on the space $(\Omega_x, \mathcal{F}_x, \mathbb{P}_x)$, which we will denote by $\mathcal{G}^{M,J,x}$.

Given a real number $M > 1$, we keep track of prime factors up to My as follows. For any integer X that is (My)-smooth, define the class $[X]$ to be the set of primes p for which $y < p < My$ and X is divisible by p to an odd power; that is, $p \in [X]$ if and only if $y < p < My$ and $p^i \,|\, X$ but $p^{i+1} \nmid X$ for some odd integer i. If X is y-smooth, we define $[X]$ to be the empty set. If X is not

(My) smooth, we pick a symbol (for probabilists, the traditional symbol is Δ) and set $[X] = \Delta$.

Now we define a random hypergraph with vertices in $\mathbb{R}^+$ by

$$\mathcal{G} := \mathcal{G}^{M,J,x} := \{[X_j] \,:\, [X_j] \neq \Delta \text{ and } \#[X_j] \geq 2\}_{1 \leq j \leq J}.$$

We remark that for a fixed x, the random hypergraphs $\mathcal{G}^{M,J,x}$ are defined simultaneously for all M and J. In case it seems strange to take $V = \mathbb{R}^+$ instead of $\mathbb{Z}^+$, it is because we will be taking scaling limits. Some easy but useful estimates are as follows.

PROPOSITION 4.3. *Fix $M > 1$ and $\eta > 0$. Let $J = \lfloor \eta J_0 \rfloor$ and let $[X_1], [X_2], \ldots$ and $\mathcal{G}$ denote the random variables on $(\Omega_x, \mathcal{F}_x, \mathbb{P}_x)$ constructed above. For any finite set S of primes, let*

$$N(S) := \#\{j : j \leq J; [X_j] = S\}.$$

(i) *For any finite set S of primes in (y, My) with $\log\log x \geq |S| \geq 2$, the number $N(S)$ has asymptotic mean*

$$\mathbb{E}_x N(S) \sim \eta \frac{y(\log y)^{|S|-1}}{\prod_{p \in S} p}. \tag{42}$$

An upper bound, with an extra factor, is valid for all S:

$$\mathbb{E}_x N(S) \leq 2^{|S|+1} \eta \frac{y(\log y)^{|S|-1}}{\prod_{p \in S} p}. \tag{43}$$

(ii) *For any set $\mathcal{W}$ of hyperedges S, let $N(\mathcal{W}) := \sum_{S \in \mathcal{W}} N(S)$ denote the total number of hyperedges in $\mathcal{W}$. Then, for any $\mathcal{W}$, $\mathbb{P}_x(N(\mathcal{W}) \geq 2) < (1/2)(\mathbb{E}_x N(\mathcal{W}))^2$.*

(iii) *For any $p \in (y, My)$, the probability that there will be a prime $q \neq p$ in (y, My) such that more than one hyperedge of $\mathcal{G}$ contains both p and q goes to zero uniformly in $M \leq M_0$, $\eta \leq \eta_0$, and $y < p \leq My$.*

Proof. The means are computed by counting the number of $a \leq x$ with $[a] = S$. The number of integers of the form $s \prod_{p \in S} p$ up to x where s is y-smooth is $\psi(x/\prod_{p \in S} p, y)$. The number of integers of this form that are divisible by q^2 for some $q \in S$ is bounded above by $\sum_{q \in S} \psi\left(\frac{x}{q \prod_{p \in S} p}, y\right)$. This is easily shown to be asymptotically negligible compared to $B_S := \psi\left(\frac{x}{\prod_{p \in S} p}, y\right)$ by (17) and Proposition 2.2, using the fact that α remains bounded away from zero; hence the number of $a \leq x$ with $[a] = S$ is asymptotically equal to B_S.

By (23), and using $\pi(y) \sim y/\log y$, we then have

$$\mathbb{E}_x N(S) \sim J \frac{\psi(x/\prod_{p\in S} p, y)}{x}$$
$$\sim \eta \frac{y(\log y)^{|S|-1}}{\prod_{p\in S} p},$$

which is (42). Using (24) instead of (23), and $\pi(y) \le 2y/\log y$ instead of $\pi(y) \sim y/\log y$, gives (43).

The second statement follows because $N(\mathcal{W})$ has a binomial distribution: If Z has binomial distribution with any parameters n and p, write $Z = \sum_{i=1}^n Y_i$ as the sum of independent Bernoulli variables to obtain

$$\mathbb{P}(Z \ge 2) \le \sum_{1\le i<j\le n} \mathbb{E} Y_i Y_j < \frac{1}{2}\mathbb{E} Z^2.$$

For the third statement, let $H(p)$ denote the event that there is some q for which more than one hyperedge arises containing p and q. Fix any primes $p_1 \ne p_2$. Let $\mathcal{W}_k$ denote the set of sets $S = \{p_1, p_2, \dots, p_k\}$ of distinct primes between y and My, and let $\mathcal{W} = \cup_{k\ge 2}\mathcal{W}_k$. By the second statement of this proposition, an upper bound for $\mathbb{P}(H(p_1))$ may be obtained by summing any upper bound for $(\mathbb{E}_x N(\mathcal{W}))^2$ as p_2 ranges over primes between y and My. We compute this by bounding $\mathbb{E}_x N(\mathcal{W}_k)$, then summing over k, squaring, and summing over p_2. Thus we begin by using (43) with $S = \{p_1, p_2, \dots, p_k\}$ to obtain

$$\mathbb{E} N(S) \le 2^{k+1}\eta y (\log y)^{k-1} \prod_{p\in S} \frac{1}{p}.$$

Summing this over all choices of $p_3, \dots, p_k$ and using (26) for the last inequality then gives

$$\mathbb{E} N(\mathcal{W}_k) \le \frac{2^{k+1}\eta\, y \log y}{p_1 p_2} \sum_{p_3<\cdots<p_k} \prod_{j=3}^k \frac{\log y}{p_j}$$
$$\le \frac{2^{k+1}\eta\, y \log y}{p_1 p_2} \frac{1}{(k-2)!} \sum_{p_3,\cdots,p_k} \prod_{j=3}^k \frac{\log y}{p_j}$$
$$\le \frac{2^{k+1}\eta\, y \log y}{p_1 p_2} \frac{1}{(k-2)!} \prod_{j=3}^k (2\log M).$$

This is valid for all $k \ge 3$, but also for $k = 2$ provided that we interpret an empty product as equal to 1 and the sum in the penultimate line as summing a single empty product. We now sum this over all integers $k \ge 2$ so that

$$\mathbb{E} N(\mathcal{W}) \le \frac{8M^4\eta y \log y}{p_1 p_2} \le \frac{8M_0^4\eta_0 \log y}{p_2}$$

since $y/p_1 < 1$. Squaring, noting that $1/p_2 < 1/y$ and $\log y < \log p_2$, we obtain a quantity bounded above by a constant multiple of

$$\frac{\log y}{y} \sum_{y<p_2\le My} \frac{\log y}{p_2}.$$

By (25) this is $O(\frac{\log y}{y})$; this completes the proof, as we only needed to show $o(1)$. □

We now define sub-hypergraphs $\mathcal{G}_{m,p}$ of the random hypergraph $\mathcal{G}$, culled so as to be tree-like and rooted at p. They are deterministic functions of the variables $X_1, \dots, X_J$, and they will bear witness to the creation of smooth products of several of the X_j. They depend on the parameters M, J, and x, which are fixed throughout the construction and suppressed in the notation.

Definition 4.4 (The sub-hypergraph $\mathcal{G}_{m,p}$ and marked set $U_{m,p}$). We define hypergraphs $\mathcal{G}_{m,p}(j)$ recursively for $m \ge 1$ and $1 \le j \le J$ as follows.

- Let $T_0(p) := \{p\}$ and $\mathcal{G}_{0,p} := \emptyset$, taking $\operatorname{supp}(\mathcal{G}_{0,p}) = \{p\}$ by convention.
- For each $m \ge 1$, define $\mathcal{G}_{m,p}(0) := \mathcal{G}_{m-1,p}$. For $j \ge 1$, define $\mathcal{G}_{m,p}(j) := \mathcal{G}_{m,p}(j-1) \cup \{[X_j]\}$ if $[X_j]$ intersects $\operatorname{supp}(\mathcal{G}_{m,p}(j-1))$ in a single element of $T_{m-1}(p)$ and $|[X_j]| \ge 2$. Otherwise, let $\mathcal{G}_{m,p}(j) := \mathcal{G}_{m,p}(j-1)$. Define $\mathcal{G}_{m,p} := \mathcal{G}_{m,p}(J)$. Define $T_m(p) := \operatorname{supp}(\mathcal{G}_{m,p}) \setminus \operatorname{supp}(\mathcal{G}_{m-1,p})$.

Let U denote the set of primes q with $y < q < My$ such that $[X_j] = q$ for some $j \le J$. Let $U_{m,p} := U \cap \operatorname{supp}(\mathcal{G}_{m,p})$. Then $(\mathcal{G}_{m,p}, U_{m,p})$ is a marked sub-hypergraph, which we will use later to witness the creation of pseudo-smooths.

Informally, $\mathcal{G}_{1,p}$ takes all hyperedges of $\mathcal{G}$ that contain p except for those creating a collision (that is, a cycle on hyperedges), using the order in which they were generated to settle collisions. Then, $\mathcal{G}_{2,p}$ starts over, taking all hyperedges containing each of the vertices added in the previous step, except for those that cause collisions. In the end, the list of hyperedges is swept through, in order, m times. The informal interpretation of $T_m(p)$ is the set of primes that first appear at distance m from p in our tree-like hypergraph; the informal interpretation of $U_{m,p}$ is the set of primes within distance m of p that appear as hyperedges of cardinality one.

LEMMA 4.5. *There are absolute constants C_ℓ such that for every η, M, x, and p,*

$$\sum_{k\ge 2} \frac{k^\ell}{(k-1)!} \sum_{p_2,\dots,p_k} \mathbb{E}_x N(p, p_2, \dots, p_k) \le \frac{\eta y}{p} C_\ell M^2 (1+\log M)^\ell,$$

where the sum is over ordered k-tuples of distinct primes in (y, My) beginning with p.

Proof. The sum $\sum_{k\geq 2} \frac{k^\ell}{(k-1)!}u^{k-1}$ may be computed exactly and is equal to $p_\ell(u)e^u - 1$, where p_ℓ is a polynomial of degree ℓ. Thus

$$\sum_{k\geq 2} \frac{k^\ell}{(k-1)!}u^{k-1} \leq c_\ell(1+u)^\ell e^u. \tag{44}$$

Using this, we may use (43) to compute the bound

$$\sum_{k\geq 2} \frac{k^\ell}{(k-1)!} \sum_{p_2,\ldots,p_k} \mathbb{E}_x N(p,p_2,\ldots,p_k) \leq \sum_{k\geq 2} \frac{k^\ell}{(k-1)!} 2^{k+1} \sum_{p_2,\ldots,p_k} \frac{\eta y}{p} \prod_{i=2}^{k} \frac{\log y}{p_i};$$

this is an upper bound so we may include terms with $p_i = p_j$. The inner sum factors as a power, and subsuming $k-1$ factors of 2 from the term 2^{k+1} into the product yields an upper bound of

$$\frac{4\eta y}{p} \sum_{k\geq 2} \frac{k^\ell}{(k-1)!} \left(\sum_{y<q<My} \frac{2\log y}{q} \right)^{k-1}.$$

By the prime number theorem, $\sum_{y<q<My}(2\log y)/q \to 2\log M$ and is never more than $2\log(2M)$, whence this bounds becomes

$$\frac{4\eta y}{p} \sum_{k\geq 2} \frac{k^\ell}{(k-1)!} \left(2\log(2M)\right)^{k-1}.$$

This is of the form (44) after taking $u = 2\log(2M)$, and applying this bound completes the proof. □

LEMMA 4.6. *For any $\eta \leq 1$ and any M, x, and p, we have the following upper bounds*:

$$\text{(i)} \qquad \mathbb{E}_x\,|\mathcal{G}_{1,p}| \leq cM^2\frac{\eta y}{p},$$

$$\text{(ii)} \qquad \mathbb{E}_x\,|\operatorname{supp}\mathcal{G}_{1,p}| \leq c'(1+\log M)M^2\frac{\eta y}{p},$$

$$\text{(iii)} \qquad \mathbb{E}_x\,|\operatorname{supp}\mathcal{G}_{1,p}|^2 \leq c''(1+\log M)^2M^4\frac{\eta y}{p},$$

$$\text{(iv)} \qquad \mathbb{E}_x\,|\operatorname{supp}\mathcal{G}_{m,p}| \leq c_m\left(1+M^3\frac{\eta y}{p}\right)^m.$$

Proof. By construction, the hypergraph $\mathcal{G}_{1,p}$ is a subset of the restriction of $\mathcal{G}$ to hyperedges containing p. Therefore,

$$\mathbb{E}_x|\mathcal{G}_{1,p}| \leq \sum_S \mathbb{E}_x N(S),$$

where the sum is over such sets S. Break down the sum by the cardinality of S. The sum over $|S| = k$ is $1/(k-1)!$ times the sum over ordered sets of

distinct primes $p = p_1, p_2, \dots, p_k$ in the range (y, My). Thus

$$\mathbb{E}_x|\mathcal{G}_{1,p}| \le \sum_{k\ge 2} \frac{1}{(k-1)!} \sum_{p_2,\dots,p_k} \mathbb{E}_x N(p, p_2, \dots, p_k).$$

Lemma 4.5 with $\ell = 0$ now gives (i). For (ii), write

$$\mathbb{E}_x|\operatorname{supp} \mathcal{G}_{1,p}| \le \sum_S |S| \mathbb{E}_x N(S)$$

and proceed as before but with $\ell = 1$.

For (iii), use

$$\mathbb{E}_x|\operatorname{supp} \mathcal{G}_{1,p}|^2 \le \sum_{S,T} |S|\,|T|\, \mathbb{E}_x N(S)N(T).$$

Observe that for $S \ne T$, the events $\{S \in \mathcal{G}_{1,p}\}$ and $\{T \in \mathcal{G}_{1,p}\}$ are negatively correlated. (Recall that two events are negatively correlated if the probability of their conjunction is at most the product of the probabilities of the events.) This is because the events $\{[X_i] = S\}$ and $\{[X_j] = T\}$ are independent, unless $i = j$, in which case they are negatively correlated. It follows that

$$\sum_{S,T} |S|\,|T|\, \mathbb{E}_x N(S)N(T) \le \sum_{S,T} |S|\,|T|\, \mathbb{E}_x N(S)\mathbb{E}_x N(T) + \sum_S |S|^2\, \mathbb{E}_x N(S)^2.$$

The first term on the right-hand side is $(\mathbb{E}_x|\operatorname{supp} \mathcal{G}_{1,p}|)^2$. Using the upper bound just established in (ii) and noting that $\eta y/p \le 1$ bounds this term by the RHS of (iii) for some constant c'''. Lemma 4.5 with $\ell = 2$ implies that the second term is bounded by a similar expression with a different constant c'''', and taking $c'' = c''' + c''''$ establishes (iii).

Finally, for (iv), induct on m. Conditional on $\mathcal{G}_{m-1,p}$, the random hypergraph $\mathcal{G}_{m,p}$ is stochastically dominated by the union of $\mathcal{G}_{m-1,p}$ with a collection of hyperedges whose conditional distribution given $\mathcal{G}_{m-1,p}$ is described as follows: for each $q \in T_{m-1}(p)$, and for each finite subset S of primes in (y, My) containing q, the hyperedge S is added independently with probability $\mathbb{P}([X_j] = S)$ which is at most $\mathbb{E}_x N(S)$. We saw in (ii) that this gives a mean of at most $O(M^3\eta y/p)$ new vertices for each q (where we have used $O(M^3)$ for an upper bound to $M^2(1 + \log M)$). By induction, is at most $c_{m-1}(1 + M^3\eta y/p)^{m-1}$ and multiplying completes the inductive step. □

Let $\mathbf{V}$ denote the vector space over $\mathbb{F}_2$ whose basis is the set of symbols

$$\{\delta_p : p \text{ is a prime and } y < p < My\}.$$

Identify each class $[X]$ with the element $\sum_{p\in[X]} \delta_p$ of $\mathbf{V}$. It is useful to think of finding "pseudo-smooth" numbers by taking products of the numbers X_j in such a way that all exponents of primes greater than y_0 are even. These pseudo-smooth numbers may be added to the list of smooth numbers, enhancing the efficiency of Schroeppel's algorithm (see the discussion after the statement of Theorem 1.2). Formally, by definition, the number of pseudo-smooths generated by time j is the difference between j and the $\mathbb{F}_2$-rank of the collection

$[X_1], \ldots, [X_j]$, made into a $\mathbb{F}_2$-vector space by using the symmetric difference operation $[X_i] \oplus [X_j]$. To count this, we count the number of j for which $[X_j]$ is in the $\oplus$-span of $[X_1], \ldots, [X_{j-1}]$, which we denote by $\langle [X_1], \ldots, [X_{j-1}] \rangle$. This includes the case where $[X_j] = \emptyset$ (y-smooth numbers), $[X_j] = [X_i] = \{p\}$ for some $i < j$ and $p < My$ (the one large prime case), as well as more complicated cases. It turns out that not much is lost if we include only one more class of cases. For each prime p in the interval (y, My), and each positive integer j, we define an event $\chi_{m,p}^{M,j}$ whose informal interpretation is that $\{p\}$ is in the span of $\{[X_1, \ldots, [X_j]\}$. A proposition immediately following the definition verifies the interpretation. The parameters x, j and M will now be fixed throughout the definition and suppressed from the notation.

Definition 4.7 (χ for general marked rooted trees).

(i) Let (G, U) be any marked hypergraph rooted at a vertex p. For $q \in \operatorname{supp}(G)$, define the height $\ell(q)$ to be the length of the longest nonbacktracking path from q to the leaves of G, or more accurately, of the tree $\mathbf{T}_p(G)$.

(ii) Define an event $\chi(q) = \chi(G, U, q)$ by recursion on $\ell(q)$. If $\ell(q) = 0$, define the event $\chi(q)$ to hold if and only $q \in U$. If $\ell(q) > 0$, let r denote the distance from p to q in $\mathbf{T}_p(G)$ and define $\chi(q)$ to hold if and only if there is some hyperedge $S \in G$ such that (a) $S \subseteq T_{r+1}(p) \cup \{q\}$ (that is, S is a hyperedge that appears first at distance $r+1$ from p, and is a "child" of q), and (b) the event $\chi(q')$ occurs for each $q' \in S$ other than q.

(iii) Finally, let $\chi(G, U)$ denote the event $\chi(G, U, p)$. (This is unambiguous because p is the root of G.)

Remarks 4.8. Note that the recursion is well founded because $\ell(q') \leq \ell(q) - 1$ for all such q'. Also note that in the recursive part of the definition, we allow S to equal $\{q\}$, in which case (b) is vacuously satisfied.

Definition 4.9 (primes witnessed in an m-neighborhood). If $\mathcal{G}_{m,p}$ is not tree-like, we define $\chi_{m,p}$ not to occur. If $\mathcal{G}_{m,p}$ is tree-like, we define $\chi_{m,p}(q) := \chi_{m,p}(\mathcal{G}_{m,p}, U_{m,p}, q)$, whence, $\chi_{m,p} := \chi(\mathcal{G}_{m,p}, U_{m,p}, p)$.

In the next proposition, $\langle [X_1], \ldots, [X_j] \rangle$ denotes the span of $\{[X_1], \ldots, [X_j]\}$ in $\mathbf{V}$.

PROPOSITION 4.10. *For any $m \geq 1$, the event $\chi_{m,p}(q)$ implies $\{q\} \in \langle [X_1], \ldots, [X_j] \rangle$. In particular,*

$$\chi_{m,p} \implies \{p\} \in \langle [X_1], \ldots, [X_j] \rangle.$$

Proof. By induction on $\ell(q) \geq 0$. If $\ell(q) = 0$, then $\chi_{m,p}(q)$ implies $[X_j] = \{q\}$ for some $j \leq J$, which immediately implies $\{q\} \in \langle [X_1], \ldots, [X_j] \rangle$. Now

suppose $\ell(q) \geq 1$. If $\chi_{m,p}(q)$ holds, let j satisfy (a) and (b) of the definition with $q = p$. For each $q' \in [X_j]$ distinct from q, $\ell(q') \leq \ell(q) - 1$, whence by induction, $\{q'\} \in \langle [X_1], \ldots, [X_j] \rangle$ for all such q'. This, along with the trivial observation that $[X_j] \in \langle [X_1], \ldots, [X_j] \rangle$, implies $\{q\} \in \langle [X_1], \ldots, [X_j] \rangle$, which completes the induction. □

The purpose of $\chi_{m,p}$ is to witness the event that $\{p\}$ is in the span of $[X_1], \ldots, [X_{j-1}]$. We wish to count this because for any J, the number of linear dependences among $\{[X_1], \ldots, [X_J]\}$ is bounded from below by

$$\#\{j \leq J : \text{ for all } p \in [X_j], \text{ the singleton } \{p\} \text{ is in the span } \langle [X_1], \ldots, [X_{j-1}] \rangle\}. \tag{45}$$

4.3. *Construction of the limit object $\mathcal{H}_m$.* An informal description of the limit object is as follows. For each $k \geq 2$, the root, ρ, gets hyperedges $\{\rho, \rho_1, \ldots, \rho_k\}$ independently, with the probability of such a hyperedge arising in a small volume element $\{\rho\} \times [\rho_1, \rho_1 + d\rho_1] \times \cdots \times [\rho_k, \rho_k + d\rho_k]$ equal to

$$\frac{d\rho_1 \cdots d\rho_k}{\rho \rho_1 \cdot \rho_k}.$$

Recursively, for m iterations, each vertex newly added in the last iteration gets new hyperedges in the same way.

Formally, the limit object is best described in terms of Poisson processes. We briefly summarize definitions and properties of these, referring the reader to [7] for further details. Given a measure space $(\mathcal{S}, \mathcal{B})$ with a σ-finite measure μ, a Poisson process with intensity μ is a collection of random variables $\{N(S) = N(S)(\omega) : S \in \mathcal{B}\}$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ satisfying the following properties:

(i) Countable additivity in S: if $\mathcal{A}$ is a collection of disjoint elements of $\mathcal{B}$, then $N\left(\bigcup_{S \in \mathcal{A}} S\right) = \sum_{S \in \mathcal{A}} N(S)$;
(ii) Counting measure: $N(S)$ takes values in the nonnegative integers;
(iii) Poisson distribution: for fixed S, the random variable $N(S)$ is distributed as a Poisson distribution with mean $\mu(S)$;
(iv) Independence: if S, T are disjoint elements of $\mathcal{B}$, then $N(S)$ and $N(T)$ are independent.

A number of constructions are available to prove the existence of such a process.

If μ is nonatomic, then with probability 1, the random counting measure N gives measure at most 1 to every point $s \in \mathcal{S}$. It follows that the random measure $N(S)$ is the sum of point masses δ_s as s ranges over some finite or countable subset of $\mathcal{S}$; we denote this set by $\operatorname{supp}(N)$, and refer to $\operatorname{supp}(N)$ as "the points of the Poisson process." The cardinality of $\operatorname{supp}(N)$ is a Poisson random variable with mean $\mu(\mathcal{S})$.

Fix a real number $M > 1$. Fix also a real $\eta > 0$. We construct a random hypergraph $\mathcal{H}_{m,\rho} = \mathcal{H}_{m,\rho}^{M,\eta}$ on a new probability space $(\Omega, \mathcal{F}, \mathbb{P})$ whose vertex set is the real interval $[1, M]$. The collection $[1, M]_k$ of subsets of $[1, M]$ of cardinality k may be identified with the sector $W_k \subseteq \mathbb{R}^k$ defined by

$$W_k := \{(\rho_1, \dots, \rho_k) \in \mathbb{R}^k : 1 \leq \rho_1 < \cdots < \rho_k \leq M\}.$$

Let $\mu_k := d\mathbf{p}/(\rho_1, \dots, \rho_k)$ denote the image under this identification of the measure whose density with respect to Lebesgue measure is $1/(\rho_1 \cdots \rho_k)$. Observe that the total mass of the measure $d\mathbf{p}/(\rho_1 \cdots \rho_k)$ is given $(\log M)^k/k!$. Now define a measure μ on the union $\bigcup_{k=1}^{\infty}[1, M]_k$ by $\mu = \sum_{k=1}^{\infty} \mu_k$. We see that μ has finite total mass:

$$||\mu|| = \sum_{k=2}^{\infty} \frac{(\log M)^k}{k!} = M - 1 - \log M.$$

Fix $\rho \in [1, M]$, and define an operation σ_ρ by $\sigma_\rho(S) = S \cup \{\rho\}$. Define the measure $\mu_k^{+\rho}$ by $\mu_k^{+\rho} = \mu_k \circ \sigma_\rho^{-1}$. In other words, $\mu_k^{+\rho}$ is the measure corresponding to "choosing a set according to μ_k" and then adding the element ρ. (Here the quotes are to remind the reader that the finite measure μ_k is not a probability measure.) Thus all the measures $\mu_k^{+\rho}$ as well as the sum $\mu^{+\rho}$ are finite measures supported on sets of finite (but not bounded) cardinality at least 2.

Let $\tau \in [1, M]$ (here τ plays the role of q/y, just as ρ plays the role of p/y). Let $\nu_\tau = \nu_\tau^{M,\eta}$ (as usual, we suppress quantities that are, for the moment, fixed) be the law of the points of a Poisson process with intensity $\eta\mu^{+\rho}/\tau$. Observe that each point of the process is a finite subset S of $[1, M]$ with $\rho \in S$. Because the intensity measure has finite mass, the law of the set of points is the law of a random finite set of hyperedges $S \subseteq [1, M]$. By nonatomicity of Lebesgue measure, we see that with probability 1, this is a tree-like hypergraph rooted at ρ, all of whose hyperedges contain ρ.

Definition 4.11 (The marked graph $(\mathcal{H}_{m,\rho}, \tilde{U}_{m,\rho})$). We now construct the random hypergraphs $\mathcal{H}_{m,\rho} = \mathcal{H}_{m,\rho}^{M,\eta}$ by recursion on m. For $m = 1$, choose $\mathcal{H}_{1,\rho}$ from the law ν_ρ. For $m \geq 1$, let $T_{m,\rho} = \operatorname{supp}(\mathcal{H}_{m,\rho}) \setminus \operatorname{supp}(\mathcal{H}_{m-1,\rho})$, taking $\operatorname{supp}(\mathcal{H}_{0,\rho}) = \{\rho\}$ by convention. For the recursion step, choose random hypergraphs $\mathcal{H}_{m,\tau}$ independently from respective laws ν_τ, as τ varies over $T_{m,\rho}$, and let $\mathcal{H}_{m+1,\rho}$ be the union of $\mathcal{H}_{m,\rho}$ with all the sets $\mathcal{H}_{m+1,\tau}$. It is again immediate that each $\mathcal{H}_{m,\rho}$ is tree-like. Finally, we define a set of marks $\tilde{U}_{m,\rho}$, by choosing each $\tau \in \operatorname{supp}(\mathcal{H}_{m,\rho})$ independently, with probability $1 - e^{-\eta/\tau}$. (We will see that $1 - e^{-\eta\tau}$ is the limiting probability for the mark $U_{m,p}$; cf. Lemma 4.16.)

Now, using Definition 4.9 once more, define events

$$\chi'_{m,\rho}(\tau) := \chi(\mathcal{H}_{m,\rho}, \tilde{U}_{m,\rho}, \tau),$$
$$\chi'_{m,\rho} := \chi(\mathcal{H}_{m,\rho}, \tilde{U}_{m,\rho}).$$

These are events on the space Ω analogous to the events $\chi_{m,p}(q)$ and $\chi_{m,p}$ defined on the space Ω_x. Denote

$$\theta_m(\rho) := \theta_m^{M,\eta}(\rho) := \mathbb{P}(\chi'_{m,\rho}).$$

4.4. *Convergence of $\mathcal{G}$ to $\mathcal{H}$, and consequently, of $\mathbb{P}_x(\chi)$ to θ.* In this subsection we prove convergence results that will be used to compute the rate of accumulation of pseudo-smooth numbers.

THEOREM 4.12. *Fix an integer $m \geq 1$ and any real $M > 1$. Then*

$$\mathbb{P}_x(\chi_{m,p}^{M,j}) = (1 + o(1))\theta_m^{M,j/J_0}(p/y) \tag{46}$$

uniformly as p varies over primes in the interval (y, My) and j/J_0 remains bounded. More generally, for any fixed $r \geq 1$ and any $p_1, \ldots, p_r$,

$$\mathbb{P}_x\left(\bigcap_{i=1}^{r} \chi_{m,p_i}^{M,j}\right) = (1 + o(1)) \prod_{i=1}^{r} \theta_m^{M,j/J_0}(p_i/y) \tag{47}$$

uniformly as $p_1, \ldots, p_r$ vary over primes in the interval (y, My).

The proof of this theorem is essentially to show that the rescaled random graph $y^{-1}\mathcal{G}_{m,p}$ converges weakly to $\mathcal{H}_{m,p/y}$. We encapsulate what we need in the following lemmas. All of these are routine Poisson convergence lemmas.

LEMMA 4.13. *As $x \to \infty$, the distance in the weak metric between the random hypergraph $\mathcal{G}_{1,p}^{M,j,x}$ and the random hypergraph $\mathcal{H}_{1,p/y}^{M,j/J_0}$ goes to zero uniformly as M and j/J_0 vary over bounded intervals and $y < p < My$.*

Proof. As a preliminary computation, let $\mathcal{G}'_{1,p}$ denote the subset of $\mathcal{G}$ of all hyperedges containing $\{p\}$. We claim that $\mathbb{P}(\mathcal{G}_{1,p} = \mathcal{G}'_{1,p}) \to 1$. Indeed, the complementary event requires that a collision occur, entailing two hyperedges both to contain $\{p\}$ and $\{q\}$ for some q. By the last part of Proposition 4.3, this probability goes to zero uniformly.

Next, let $\Xi = (\tau_1, \tau'_1] \times \cdots \times (\tau_k, \tau'_k]$ be any rectangular subset of the sector W_k, and let Ξ_x denote the set of sets, S, of k primes, each between y and My, such that $y^{-1}S \in \Xi$. As in Proposition 4.3, let $N(\sigma_p(\Xi_x))$ denote the number of $j \leq J$ such that $[X_j] \in \sigma_p(\Xi_x)$. Using (42), we estimate

$$\begin{aligned}\mathbb{E}_x N(\sigma_p(\Xi_x)) &= \sum_{S \in \sigma_p(\Xi_x)} \mathbb{E}_x N(S) \\ &\sim \sum_{S \in \sigma_p(\Xi_x)} \eta \frac{y(\log y)^k}{p \prod_{q \in S} q}.\end{aligned}$$

Factoring the sum of products gives the equivalent expression

$$\mathbb{E}_x N(\sigma_p(\Xi_x)) \sim \eta \frac{y}{p} \prod_{i=1}^{k} \sum_{\tau_i y < q \le \tau_i' y} \frac{\log y}{q}.$$

By the prime number theorem, this converges to $\nu_{p/y}(\Xi)$.

Finally, let us see that $y^{-1}\mathcal{G}_{1,p}$ converges to a Poisson process with intensity ν_ρ where $\rho = p/y$; by construction, this is the distribution of $\mathcal{H}_{1,\rho}$, and therefore this will complete the proof of the lemma. Recalling the remark at the end of Section 4.1, we need to show (i) tightness and (ii) that for any disjoint sets $\Xi^{(1)}, \dots, \Xi^{(k)}$, the respective numbers $N^{(i)}$ of hyperedges in $y^{-1}\mathcal{G}_{1,p}$ in $\Xi^{(i)}$ converge in disribution to independent Poissons with means $\nu_\rho(\Xi_i)$. It suffices to prove these for $\mathcal{G}'_{1,p}$ in place of $\mathcal{G}_{1,p}$ because we have seen these are equal with probability $1 - o(1)$.

We have already verified that the means are $\nu_\rho(\Xi^{(i)})$ when $\Xi^{(i)}$ are rectangles, which implies the same result for all measurable Ξ. To obtain the joint Poisson distribution, it is easiest to Poissonize. Replace $\mathcal{G}'_{1,p}$ by $\mathcal{G}''_{1,p}$, defined identically to $\mathcal{G}'_{1,p}$ except with J replaced by a Poisson variable J' of mean J. For this random graph, the numbers $(N^{(i)})''$ of hyperedges of $\mathcal{G}''_{1,p}$ in the rescaled $\Xi^{(i)}$ are exactly independent Poissons with the given means. The key observation is that

$$\mathbb{P}_x(\mathcal{G}'_{1,p} \neq \mathcal{G}''_{1,p}) = O(J_0^{-1/2}).$$

To see this, note that $\mathbb{E}_x|J' - J| = O(\sqrt{J_0})$. Therefore,

(48)
$$\mathbb{P}_x(\mathcal{G}'_{1,p} \neq \mathcal{G}''_{1,p}) = O\left(\sqrt{J_0}\,\mathbb{P}_x(p \in [X_1])\right) = O\left(J_0^{-1/2}\,\mathbb{E}_x|\mathcal{G}'_{1,p}|\right) = O\left(J_0^{-1/2}\right)$$

by Lemma 4.6. Finally, tightness also follows from Lemma 4.6. □

LEMMA 4.14. *As $x \to \infty$, the distance in the weak metric between the n-tuple of random hypergraphs*

$$y^{-1}\left(\mathcal{G}_{1,p_i}^{M,j,x}\right)_{1 \le i \le n}$$

and the product of the laws of the hypergraphs $\mathcal{H}_{1,p_i/y}^{M,j/J_0}$ goes to zero uniformly as M and j/J_0 vary over bounded intervals and $y < p_i < My$.

Proof. This is the same proof with only one difference, as follows. To check that $\mathcal{G}_{1,p_i} = \mathcal{G}'_{1,p_i}$ with probability tending to 1, one observes that (3) of Proposition 4.3 holds simultaneously for $p_1, \dots, p_n$. All else is the same, once one observes that Poissonization gives (48) simultaneously for all $p_1, \dots, p_n$. □

LEMMA 4.15. *As $x \to \infty$, the distance in the weak metric between the random hypergraph $y^{-1}\mathcal{G}_{m,p}^{M,j,x}$ and the random hypergraph $\mathcal{H}_{m,p/y}^{M,j/J_0}$ goes to zero uniformly as M and j/J_0 vary over bounded intervals and $y < p < My$. Similarly, the distance between the law of the random n-tuple $y^{-1}(\mathcal{G}_{m,p_i}^{M,j,x})_{1 \le i \le n}$*

and the product of the laws of $\mathcal{H}_{m,p_i/y}^{M,j/J_0}$ *goes to zero with the same uniformity in* $M, j/J_0$ *and* $\{p_i\}$.

Proof. We induct on m. For $m = 1$, this was shown in Lemma 4.13. Now let $m \geq 2$, and assume for induction that the result holds for $m-1$. If $\mathcal{G}_{m,p}$ is tree-like, let $r := |T_1(p)|$ and let $G_1, \dots, G_r$ denote the subtrees of $\mathbf{T}_p(\mathcal{G}_{m,p})$ from the vertices $q_1, \dots, q_r$ of $T_1(p)$. Let $\mathcal{G}(1), \dots, \mathcal{G}(r)$ denote the corresponding hypergraphs; that is, $\mathcal{G}(i)$ is the hypergraph rooted at q_i whose hyperedges are those of $\mathcal{G}_{m,p}$ whose support is a subset of the vertices of G_i. We will show that the joint conditional distribution of $y^{-1}(\mathcal{G}(1), \dots, \mathcal{G}(r))$ given $\mathcal{G}_{1,p}$ converges to the product of the laws of $\mathcal{H}_{m-1,q_i/y}$. By the recursive construction of $\mathcal{H}_{m,p/y}$ and the fact that $\mathcal{G}_{m,p}$ is tree-like with probability approaching 1, this will complete the proof of the lemma.

Consider the hypergraph $\mathcal{G}'_{m-1,q_i}$. If this is tree-like, let H_i be the subtree obtained by removing the unique hyperdege containing p and q_i, and restricting to the connected component rooted at q_i. If these are disjoint for $1 \leq i \leq r$, then $\mathcal{G}(i) = H_i$ for each i. The probability that all the hypergraphs $\mathcal{G}'_{m-1,q_i}$ are tree-like is asymptotically 1. The probability of a collision is bounded above by

$$\sum_{y<q<My} \sum_{i,j=1}^{r} \mathbb{P}_x \left(q \in \operatorname{supp}(\mathcal{G}'_{m-1,q_i}) \cap \operatorname{supp}(\mathcal{G}'_{m-1,q_j}) \right).$$

The probability that $q \in \operatorname{supp}(\mathcal{G}'_{m-1,q_j})$, conditional on $|\operatorname{supp} \mathcal{G}'_{m-1,q_j}|$, is at most a constant multiple of $|\operatorname{supp} \mathcal{G}'_{m-1,q_j}|/\pi(y)$. This is true as well for q_i, and the two events are independent. Therefore, the probability of a collision is

$$O\left((\mathbb{E}_x r^2) \frac{(\mathbb{E}_x |\operatorname{supp} \mathcal{G}_{m-1,q_i}|)^2}{\pi(y)} \right).$$

By Lemma 4.6, we obtain the upper bound $O(1/\pi(y))$.

Next, we claim that the conditional distribution of H_i given $\mathcal{G}'_{1,p}$ is asymptotically equal to the unconditional distribution of $\mathcal{G}'_{m-1,q_i}$. Indeed, $\mathcal{G}'_{1,p}$ is measurable with respect to the σ-field generated by the events $\{S \in \mathcal{G} : p \in S\}$. This is independent of the events $\{S \in \mathcal{G} : p \notin S\}$, so conditional on $\mathcal{G}'_{1,p}$, H_i has the distribution of $\mathcal{G}''_{m-1,q_i}$, where the double prime means that all hyperedges containing p were excluded at every step of the construction. We already know that $\mathcal{G}''_{m-1,q-1}$ is asymptotically distributed as $\mathcal{G}'_{m-1,q-1}$, verifying the claim. Moreover, the same argument shows that the joint conditional law of $H_1, \dots, H_r)$ given $\mathcal{G}'_{1,p}$ is asymptotically the product of the laws for each $i \leq r$.

Finally, by the induction hypothesis, the unconditional distribution of $\mathcal{G}'_{m_i,q_i}$ is asymptotically that of $\mathcal{H}_{m-1,q_i/y}$. Therefore, since with probability approaching 1 all the graphs $\mathcal{G}'_{m-1,q_i}$ are tree-like and there are no collisions, we have shown what we need. □

LEMMA 4.16. *As $x \to \infty$, the distance in the weak metric between the random marked hypergraph $y^{-1}(\mathcal{G}_{m,p}^{M,j,x}, U_{m,p})$ and the random marked hypergraph $(\mathcal{H}_{m,p/y}^{M,j/J_0}, \tilde{U}_{m,p/y})$ goes to zero uniformly as M and j/J_0 vary over bounded intervals and $y < p < My$. More generally, the distance between an n-tuple of marked graphs*

$$y^{-1}\left(\mathcal{G}_{m,p_i}^{M,j,x}, U_{m,p_i}\right)_{1\le i\le n}$$

and the product of the laws of the random marked hypergraphs $(\mathcal{H}_{m,p_i/y}^{M,j/J_0}, \tilde{U}_{m,p_i/y})$ goes to zero uniformly as M and j/J_0 vary over bounded intervals and $y < p_1, \dots, p_n < My$.

Proof. Observe that the conditional probabilities of $q \in U_{m,p}$ given $\mathcal{G}_{m,p}$ are independent and given by $1 - e^{-\eta y/q}$ as q varies over $\operatorname{supp}(\mathcal{G}_{m,p})$. This is true since, in the limit ($x, y \to \infty$ and $J = \eta x \pi(y)/\psi(x,y)$), the events $|\{j : [X_j] = \{q_i\}, j = 1, \dots, J\}|$ for fixed $q_1, q_2, \dots, q_r$ are independent Poisson random variables with mean $\sim \eta y/q_i$. And once it is known, in the limit, that the events $\{q \in U_{m,p}\}$ given $\mathcal{G}_{m,p}$, with q running over $\operatorname{supp}(\mathcal{G}_{m,p})$ are independent with probability $1 - e^{-\eta y/q}$, then the first part of the lemma is proven; the second part is analogous. □

Proof of Theorem 4.12. Begin with (46). For any marked graph (G, U), $\chi(G, U)$ depends only on the marked hypergraph structure of (G, U) and not the names of the vertices. Because the topology on graph structure is discrete, χ is continuous. The weak topology on measure is characterized by convergence of integrals of bounded continuous functions, so (46) follows from the first conclusion of Lemma 4.16. For any fixed bounded continuous function, such as χ, the difference in the integrals is bounded as a function of the distance bewteen the measures, whence the uniform convergence in Lemma 4.16 transfers to the required uniform convergence in (46). The proof of (47) is identical, using the n-tuple convergence in Lemma 4.16 in place of convergence of the single marked hypergraph. □

4.5. *Computation of θ.* We begin by computing $\theta_m(\rho)$. Recall the definition of the functions $\gamma_{m,M}(u)$ in (39).

LEMMA 4.17.

$$\theta_m^{M,\eta}(\rho) = 1 - e^{-\gamma_{m,M}(\eta)/\rho}.$$

Proof. The quantities M and η will be fixed throughout the proof, so we write θ_m for $\theta_m^{M,\eta}$. The proof is by induction on m. By definition, $1 - \theta_0(\rho)$ is the probability that $\rho \notin \tilde{U}_{m,\rho}$, which is $e^{-\eta/\rho}$ by construction. This establishes the result for $m = 0$.

Now suppose that $m \ge 1$. The set of hyperedges $S \in \mathcal{H}_{1,\rho}$ is, by construction, a Poisson process with intensity ν_ρ. The complement of $\chi_{m,\rho}$ is the

intersection of $\rho \notin \tilde{U}_{m,\rho}$ with the event that for all hyperedges $S \in \mathcal{H}_{1,\rho}$ of cardinality between 2 and k, there is some $\tau \in S \setminus \{\rho\}$ with $\chi_{m-1,\tau}$ not occurring. We have, by induction,

$$1 - \theta_{m+1}(\rho) = e^{-\eta/\rho}\, \mathbb{E}\Big[\prod_{S \in \mathcal{H}_{1,\rho}} \Big(1 - \prod_{\tau \in S\setminus\{\rho\}} \theta_m(\tau)\Big)\Big] \tag{49}$$
$$= e^{-\eta/\rho}\, \mathbb{E}\Big[\prod_{S \in \mathcal{H}_{1,\rho}} \Big(1 - \prod_{\tau \in S\setminus\{\rho\}} \theta_m(\tau)\Big)\Big],$$

where the first product is over hyperedges of cardinality up to k and the product over $\tau \in S \setminus \{\rho\}$ is taken to be 1 if $S = \{\rho\}$. If $f : \Xi \to [0,1]$ is any function on a space Ξ on which is defined a Poisson process with intensity ν, then the expected product of f at points of the Poisson process is given by

$$\exp\Big[\int (f(\xi) - 1)\, d\nu(\xi)\Big].$$

Applying this to (49) with $\nu = \nu_\rho$ and $f(S) = 1 - \prod_{\tau \in S\setminus\{\rho\}} \theta_m(\tau)$ gives

$$\log(1 - \theta_{m+1}(\rho)) = -\frac{\eta}{\rho} - \int \prod_{\tau \in S\setminus\{\rho\}} \theta_m(\tau)\, d\nu(S).$$

Break up the integral according to $|S|$. Recall that for $k \geq 2$, the law of $S \setminus \{\rho\}$ on $\{|S| = k\}$ is $\eta\mu_{k-1}/\rho$. We may incorporate $-\eta/\rho$ as the $k = 1$ term if we define μ_0 to be a point mass of 1 at the empty set and the empty product to be 1. These substitutions yield

$$\log(1 - \theta_{m+1}(\rho)) = -\frac{\eta}{\rho} \sum_{k=1}^{\infty} \int \prod_{\tau \in S\setminus\{\rho\}} \theta_m(\tau)\, d\mu_{k-1}(S).$$

Observe that μ_k is $1/k!$ times a product measure. Therefore the integral of the product factors, yielding (with $j := k - 1$)

$$\log(1 - \theta_{m+1}(\rho)) = -\frac{\eta}{\rho} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\int_1^M \theta_m(\tau) \frac{d\tau}{\tau} \right)^j = -\frac{\eta}{\rho} \exp\left(\int_1^M \theta_m(\tau) \frac{d\tau}{\tau} \right).$$

Using the induction hypothesis again we substitute $1 - e^{-\gamma_{m,M}(\eta)/\tau}$ for $\theta_m(\tau)$ to arrive at

$$\log(1 - \theta_{m+1}(\rho)) = -\frac{\eta}{\rho} \exp\left(\int_1^M \left(1 - e^{-\gamma_{m,M}(\eta)/\tau}\right) \frac{d\tau}{\tau} \right).$$

Changing variables to $t = 1/\tau$ so that $dt/t = -d\tau/\tau$ yields

$$\log(1 - \theta_{m+1}(\rho)) = -\frac{\eta}{\rho} \exp\left(\int_{1/M}^1 \frac{1 - e^{-t\gamma_{m,M}(\eta)}}{t}\, dt \right)$$
$$= -\frac{\eta}{\rho} A_M(\gamma_{m,M}(\eta)).$$

The right-hand side is equal to $-(1/\rho)\gamma_{m+1,M}(\eta)$, completing the induction. □

LEMMA 4.18. *Fix any $\eta > \eta_* = e^{-\gamma}$. Then*

$$\theta_m^{M,\eta}(\rho) \to 1$$

uniformly over ρ in any bounded interval $[1, L]$ as $m, M \to \infty$.

Proof. The function $z/\exp(A(z))$ is the real analytic function

$$\exp\left(\int_1^z \frac{e^{-u}}{u}\,du - \int_0^1 \frac{1-e^{-u}}{u}\,du\right) = \exp(-\gamma - \Gamma(0,z)),$$

where $\Gamma(0,z) := \int_z^\infty e^{-t}\frac{dt}{t}$. By (28), which evidently increases to η_* as $z \uparrow \infty$. It follows that for $\eta > \eta_*$, if we choose any positive $\delta < (\eta/\eta_*) - 1$, then

$$\frac{\eta}{1+\delta} > \eta_* > \frac{z}{e^{A(z)}},$$

which implies that

$$\eta e^{A(z)} > (1+\delta)z$$

for all $z > 0$. Applying this to (39) with $z = \gamma_{m,\infty,\infty}(u)$ leads to

$$\gamma_{m+1,\infty}(\eta) > (1+\delta)\gamma_{m,\infty}(\eta)$$

which, in turn, leads inductively to

$$\gamma_{m,\infty}(\eta) > \eta_*(1+\delta)^{m-1}.$$

Since γ is increasing in all its arguments, this is true for all greater η as well.

Now, given $L, \varepsilon > 0$, choose m sufficiently large so that $\gamma_{m,\infty}(\eta) > L\log(1/\varepsilon)$. The function γ is continuous in M at infinity, so we may choose M such that $\gamma_{m,M}(\eta) > L\log(1/\varepsilon)$. It follows from Lemma 4.17 that

$$\theta_m^{M,\eta}(\rho) = 1 - e^{-\gamma_{m,M}(\eta)/\rho} > 1 - e^{-\log(1/\varepsilon)} = 1 - \varepsilon$$

for $1 \le \rho \le L$, proving the lemma. □

4.6. *Proof of main theorems.*

Proof of Theorem 1.2. Fix $\varepsilon > 0$. The first step is to use Lemma 4.18 to pick m and M such that

$$\theta_m^{M,\eta_*+\varepsilon}(\rho) > \frac{3}{4} \quad \text{for all} \quad 1 \le \rho \le L := \exp\left(\frac{3}{\varepsilon}\right).$$

Take M to be larger if necessary so that we may assume $M \ge L$. We deduce from the last displayed estimate with $\rho = p/y$ and from Theorem 4.12 that, for any prime p in the interval (y, My) and for x sufficiently large, we have

$$\mathbb{P}_x\left(\chi_{m,p}^{M,(\eta_*+\varepsilon)J_0}\right) > \frac{3}{4}.$$

Now let Y be the number of j in the interval $I := [(\eta+\varepsilon)J_0, (\eta+2\varepsilon)J_0]$ such that $[X_j] = \{p\}$ for some prime p with $y < p < My$ and $\chi_{m,p}^{M,j-1}$ holds. Write $Y = \sum_{j\in I} Y_j$, where Y_j is 1 if $[X_j] = \{p\}$ for some prime $y < p < My$ and $\chi_{m,p}^{M,j-1}$ holds, and zero otherwise. We compute a lower bound on $\mathbb{E}_x Y$ as

follows. The event $\chi_{m,p}^{M,j-1}$ is independent of the event $[X_j] = \{p\}$. By (23) and the definition of J_0, we have $\psi(x/p, y)/\psi(x, y) \sim (\log y)/p$. Hence,

$$\begin{aligned}\mathbb{E}_x Y &= \sum_{j\in I} \sum_{y<p<My} \mathbb{P}([X_j] = \{p\})\mathbb{P}(\chi_{m,p}^{M,j-1}) \\ &= \sum_{j\in I} \sum_{y<p<My} \frac{\psi(x/p, y)}{x} \mathbb{P}(\chi_{m,p}^{M,j-1}) \\ &\geq \frac{1}{2} \sum_{j\in I} \sum_{y<p<My} \frac{\pi(y)}{J_0} \frac{\log y}{p}\end{aligned}$$

for x sufficiently large. By the prime number theorem,

$$\sum_{y<p<My} (\log y)/p \sim \log M \geq \log L = 3\varepsilon^{-1}.$$

The outer sum has at least εJ_0 terms; hence,

$$\mathbb{E}_x Y \geq \frac{1}{2}(\varepsilon J_0)\frac{\pi(y)}{J_0}(3\varepsilon^{-1}) = \frac{3}{2}\pi(y). \tag{50}$$

In Lemma 4.19 below, we will prove the second moment bound

$$\operatorname{Cov}(Y_i, Y_j) = o\left(\mathbb{E}_x Y_i \mathbb{E}_x Y_j\right).$$

Using, this lemma,

$$\begin{aligned}\operatorname{Var}(Y) &= \sum_{i,j\in I} \operatorname{Cov}(Y_i, Y_j) \\ &\leq \mathbb{E}_x Y + 2 \sum_{i,j\in I, i<j} \operatorname{Cov}(Y_i, Y_j) \\ &= o(\mathbb{E}_x Y)^2.\end{aligned}$$

Together with (50), this implies that $\mathbb{P}_x(Y > \pi(y)) \to 1$. Recall from (45) that this implies more than $\pi(y)$ linear dependences among the classes $[X_j]$ with $j \leq (\eta_* + 2\varepsilon)J_0$. Since $\varepsilon > 0$ was arbitrary, this completes the proof of the theorem, modulo the lemma. □

Proof of Theorem 4.1. In the previous section, we chose M to be absurdly large, which allowed us to use only those j in the interval $[(\eta_*+\varepsilon)J_0, (\eta_*+2\varepsilon)J_0]$. We can get much more reasonable values of m and M if we are willing to let η be a little bigger and to use all the values of j up to ηJ. The computations are in fact no harder (although the required convergence lemmas did involve more work in the previous sections).

Fix η, m, and M satisfying the inequality in the hypothesis of the theorem. Let

$$Z := \sum_{j=1}^{J} Z_j := \#\left\{j \leq J : \chi_{m,p}^{M,j-1} \text{ occurs for all } p \in [X_j]\right\}.$$

Again, Lemma 4.19 implies $\mathrm{Var}\,(Z) = o(\pi(y)^2)$. If we are able to show

$$\liminf_{x\to\infty} \frac{\mathbb{E}_x Z}{\pi(y)} > 1, \tag{51}$$

then we will have $\mathbb{P}_x(Z > \pi(y)) \to 1$, which will imply more than $\pi(y)$ linear dependences, thus establishing the theorem.

To prove (51), break down $\mathbb{E}Z_j$ according to the value of $[X_j]$ and using independence of X_j from $\chi_{m,p}^{M,j-1}$. This gives

$$\begin{aligned}\mathbb{E}_x Z_j &= \sum_S \mathbb{P}_x([X_j] = S) \prod_{p\in S} \mathbb{P}_x(\chi_{m,p}^{M,j-1}) \\ &= \sum_S \frac{\psi(x/\prod_{p\in S} p, y)}{x} \prod_{p\in S} \mathbb{P}_x(\chi_{m,p}^{M,j-1}) \\ &\sim \sum_S \frac{(\log y)^{|S|}}{\prod_{p\in S} p} \frac{\psi(x,y)}{x} \prod_{p\in S} \theta_m^{M,j/J_0}(p/y).\end{aligned} \tag{52}$$

The final asymptotic equality uses equation (23) and formula (47) of Theorem 4.12 to identify the limit when summing over S of bounded cardinality. This implies half the the asymptotic equality, i.e., $\liminf[\mathbb{E}_x Z_j/\text{RHS of (52)}] \geq 1$, which is all we will need for (51). However, since it may be useful in the future, we point out that (24) may be used to show that the contribution from $|S| > K$ is $o(\pi(y))$ as $K \to \infty$, establishing (52).

Continuing, we use the identity $\psi(x,y)/x = \pi(y)/J_0$, factor out this term, and rewrite the summand as a product:

$$\mathbb{E}Z_j = \frac{\pi(y)}{J_0} \sum_S \prod_{p\in S} \left(\frac{\log y}{p} \theta_m^{M,j/J_0}(p/y) \right).$$

Let B be any set and $\{z_p : p \in B\}$ be any positive real numbers with finite sum. Let $\mathcal{B}$ denote the set of finite subsets of B. Then

$$\sum_{S\in\mathcal{B}} \prod_{p\in S} z_p = \prod_{p\in S} (1 + z_p) \to \exp\left(\sum_{p\in B} z_p \right)$$

as $\max_{p\in B} z_p \to 0$. Using this identity, we obtain

$$\begin{aligned}\mathbb{E}_x Z_j &\sim \frac{\pi(y)}{J_0} \exp\left(\frac{1}{y} \sum_{y<p<My} \frac{\log y}{p/y} \theta_m^{M,j/J_0}(p/y) \right) \\ &\sim \frac{\pi(y)}{J_0} \exp\left(\int_1^M \frac{1}{t} \theta_m^{M,j/J_0}(t)\, dt \right)\end{aligned}$$

by the prime number theorem. The asymptotic equivalence is uniform in $j \le \eta J_0$. Summing from $j = 1$ to ηJ_0 now gives

$$\frac{\mathbb{E}_x Z}{\pi(y)} \sim \int_0^\eta \exp\left(\int_1^M \frac{1}{t}\theta_m^{M,u}(t)\,dt\right)\,du$$
$$= \int_0^\eta \frac{\gamma_{m+1}(u)}{u}\,du.$$

By the hypothesized inequality, the right-hand side is greater than 1, which establishes (51) and completes the proof of the theorem. □

LEMMA 4.19. *Fix a finite real $M > 1$ and $\eta > 0$ and an integer $m \ge 1$. Then*

$$\mathrm{Cov}\,(Z_i, Z_j) = o\,(\mathbb{E}_x Z_i \mathbb{E}_x Z_j) = o\left(\frac{\pi(y)^2}{J_0^2}\right)$$

for all $1 \le i < j \le \eta J_0$. The same is true with $\mathrm{Cov}\,(Y_i, Y_j)$ in place of $\mathrm{Cov}\,(Z_i, Z_j)$.

Proof. Both arguments are the same, so we prove this just for $\mathrm{Cov}\,(Z_i, Z_j)$. It is equivalent to show that

$$\mathbb{E}_x(Z_i \cdot Z_j) \sim (\mathbb{E}_x Z_i) \cdot (\mathbb{E}_x Z_j).$$

Conditioning on $[X_i]$ and $[X_j]$, we see that this is the expectation of

$$\mathbb{E}_x(Z_i|[X_i],[X_j]) \cdot \mathbb{E}_x(Z_j|[X_i],[X_j]).$$

The sets $[X_i]$ and $[X_j]$ are disjoint with probability going to 1, so it suffices to show that $\mathbb{E}_x(Z_i|[X_i],[X_j])$ and $\mathbb{E}_x(Z_j|[X_i],[X_j])$ are asymptotically independent when $[X_i]$ and $[X_j]$ are disjoint. We have seen in Lemma 4.15 that the collection of hypergraphs $\mathcal{G}_{m,p}^{M,i-1,x}$ for $p \in [X_i]$ and $\mathcal{G}_{m,p}^{M,j-1,x}$ for $p \in [X_j]$ are disjoint and tree-like with probability going to 1, and asymptotically independent. By Lemma 4.16, the same is true of the marked hypergraphs. Since Z_i is a bounded function of $[X_i]$ and the marked hypergraphs $(\mathcal{G}_{m,p}^{M,i-1,x}, U_{m,p}^{M,i-1,x})$ for $p \in [X_i]$, and likewise for $[Z_j]$, we have the desired conditional independence. □

5. **Implications for factoring algorithms**

In factoring algorithms we need to find a linear dependence mod 2 in our matrix of exponents. We expect that the best algorithms known, due to Wiedemann or Lanczos (see Section 6.1.3 of [4]), take time

$$\sim C\frac{y^2}{\log y \log\log y}$$

for a positive constant C when we use the primes up to y in our "factor base". If we were to take $y = y_0$, then this number would be far larger than J_0 and so would dominate the running time of the algorithm. Hence, to optimize, we

select $y = y_1$, which is far smaller, chosen to equalize the running times of the two main parts of the algorithm, so that

$$c\frac{\pi(y)}{\Psi(x,y)/x} \sim \frac{y^2}{\log y \log\log y} \tag{53}$$

for an appropriate constant $c > 0$. One can show that one then has

$$y_1 = y_0^{1-(1+o(1))/\log\log x}$$

with expected running time

$$J_0\ y_0^{(1+o(1))/(\log\log x)^2}$$

(see [5]).

The proofs in the previous section work as well for y_1 as for y_0. In particular, we can determine the speed-up for various choices of the parameters. We always take $m = \infty$; instead of the many large prime variation, we consider the k-large prime variations with $1 \le k \le 5$, as has been done in algorithms that have been implemented (see [5] for more details).

k	$M = \infty$	$M = 100$	$M = 10$
0	1	1	1
1	.7499	.7517	.7677
2	.6415	.6448	.6745
3	.5962	.6011	.6422
4	.5764	.5823	.6324
5	.567	.575	.630

Table 1. The value of η such that there are $\sim \pi(y)$ pseudosmooths amongst the a_j with $j \le \eta\pi(y)x/\Psi(x,y)$.

So what effect will this reduction in the number of a_j examined have in the actual running time? Suppose that we replace c in (53) by ηc and determine that the new running time is given by (53), after solving (53) to determine $y = y_\eta$.

Now finding this solution is tantamount to finding a solution to $h(u_\eta) = \log(c\eta\log\log y)$ where $h(u) := \frac{1}{u}\log x + \log\rho(u)$. We have $h'(u) = -1 - (1 + o(1))/\log u)$, and so $u_1 - u_\eta = \log\eta(1 - (1 + o(1))/\log u)$. Our running time therefore changes by a factor of

$$\begin{aligned}\sim x^{\frac{2}{u_\eta}-\frac{2}{u_1}} &= \exp\left(\frac{2(u_1-u_\eta)\log x}{u_1u_\eta}\right) = \exp\left(\frac{2\log\eta\log x}{u_1^2}\left(1-\frac{1+o(1)}{\log u}\right)\right)\\ &= \exp\left(\log\eta(\log\log x + \log\log\log x - \log 2 - 4 + o(1))\right)\\ &= \left(\frac{2e^4+o(1)}{\log x\log\log x}\right)^{\log(1/\eta)}\end{aligned}$$

since $\log^2 y_1 = \log^2 L(x)\left(1 + \frac{\log\log\log x - \log 2 - 4 + o(1)}{\log\log x}\right)$.

Data on the effect of large prime variations that has been gathered from running factoring algorithms seems rather different from what we have obtained here. One reason for this is that, in our analysis, the variations in M and k simply affect the number of a_j being considered, whereas in reality these affect not only the number of a_j being considered but also several other important quantities — for instance, the amount of sieving that needs to be done, and also the amount of data that needs to be "swapped." (Typically one saves the a_j with several large prime factors to the disk, or somewhere else suitable for a lot of data.) It is an interesting problem to try to properly analyze the construction of programs so as to incorporate the results that we have obtained and to get predictions that would help the choice of parameters in computer algorithms. In discussion, David Moulton noted that a slight variant of Pomerance's problem allows us to fully analyze a slight variant of Dixon's random squares algorithm. Let us suppose that $n = pq$, the product of two primes. There are $\phi(n)/4$ squares mod n up to n, and in Dixon's algorithm the a_i are randomly chosen amongst these numbers. One usually makes the assumption that this is not much different from choosing random numbers up to n for the purpose of finding square products — we wish to make no assumptions.

Suppose that we know also $u, v \pmod n$ for which $(u/p) = -(v/p) = 1$ and $(u/q) = -(v/q) = -1$. The sets A, uA, vA, uvA, where A is the set of squares mod n, partition the reduced residues mod n. Hence each time we choose b_i randomly (as at the beginning of the introduction) we also select $\sigma_{u,i}, \sigma_{v,i}$ independently to have equal probability of being either 0 or 1, and then we let a_i be the least residue of $b_i^2 u^{\sigma_{u,i}} v^{\sigma_{v,i}} \pmod n$. Then each reduced residue mod n is chosen with equal probability. Our analysis of Pomerance's problem may be applied to these a_i. (It is easy to see that restricting to numbers relatively prime to n changes the algorithm with probability $o(1)$.) The upper bound in Theorem 1.2 implies that we obtain many independent congruences over $\mathcal{F}_2$ once $J > (e^{-\gamma} + \varepsilon)J_0$. In particular, we obtain at least three, which is good enough to imply that some sub-product of these three has an even power of both u and v. The result is a congruence of two perfect squares modulo n, which has at least a $1/2$ probability of factoring n and which was the goal of the original Dixon algorithm.

This shows that the upper bound of $e^{-\gamma}J_0$ for for Pomerance's approximation to Dixon's algorithm is an upper bound for Dixon's actual algorithm together with an oracle to produce u and v. Unfortunately the existence of such an oracle is equivlent to already knowing how to factor n. This problem may be surmounted as follows. Choose random numbers $u_1, \dots, u_r$ in place of u and v. It is easy to test and reject numbers u with $(u/n) = 1$, so we may assume that u_j are uniform on numbers with Jacobi symbol -1. Modify the

previous algorithm by taking

$$X_i = b_i^2 \prod_{j=1}^{r} u_j^{t_{ij}} \pmod{n},$$

where t_{ij} are independent Bernoulli (1/2) random variables (fair coin-flips). The algorithm for Pomerance's problem will eventually find $r+1$ independent congruences. When it does so, some sub-product yields a congruence of squares modulo n, and we are done. The question is, how long could this take? Our analysis of Pomerance's algorithm implies that time $(e^{-\gamma}+\varepsilon)J_0$ suffices, as long as the numbers X_i are uniform on $\{1,\dots,n\}$. Again it is good enough to be uniform on the multiplicative group $G := (\mathbb{Z}/n\mathbb{Z})^*$. This will be the case as long as $u_1,\dots,u_r$ generate all of G/G^2, which is isomorphic to $(\mathbb{Z}/2\mathbb{Z})^2$. Each u_j has probability 1/2 of being in each of the two cosets for which $(u/n) = -1$, and they fail to generate G/G^2 only if all are in the same coset, which has probability 2^{1-r}. Letting r grow sufficiently slowly, we see that this modification of Dixon's algorithm terminates with success in time $(e^{-\gamma}+\varepsilon)J_0$ with probability $1-o(1)$. Thus our results for Pomerance's problem is in fact a bound for the original Dixon algorithm, provided we keep track of a few more things along the way.

References

[1] M. Abramowitz and I. A. Stegun (eds.), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover Publications, New York, 1965.

[2] J. P. Buhler, H. W. Lenstra, Jr., and C. Pomerance, Factoring integers with the number field sieve, in *The Development of the Number Field Sieve*, *Lecture Notes in Math.* **1554**, Springer-Verlag, New York, 1993, pp. 50–94. MR 1321221. Zbl 0806.11067. http://dx.doi.org/10.1007/BFb0091539.

[3] N. J. Calkin, Dependent sets of constant weight binary vectors, *Combin. Probab. Comput.* **6** (1997), 263–271. MR 1464565. Zbl 0887.60015. http://dx.doi.org/10.1017/S0963548397003040.

[4] R. Crandall and C. Pomerance, *Prime Numbers*; *A Computational Perspective*, second ed., Springer-Verlag, New York, 2005. MR 2156291. Zbl 1088.11001.

[5] E. Croot, A. Granville, R. Pemantle, and P. Tetali, Running time predictions for factoring algorithms, in *Algorithmic Number Theory*, *Lecture Notes in Comput. Sci.* **5011**, Springer-Verlag, New York, 2008, pp. 1–36. MR 2467835. Zbl 1205.11132. http://dx.doi.org/10.1007/978-3-540-79456-1_1.

[6] J. D. Dixon, Asymptotically fast factorization of integers, *Math. Comp.* **36** (1981), 255–260. MR 0595059. Zbl 0452.10010. http://dx.doi.org/10.2307/2007743.

[7] R. Durrett, *Probability. Theory and Examples*, *The Wadsworth & Brooks/Cole Statistics/Probability Series*, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1991. MR 1068527. Zbl 0709.60002.

[8] E. Friedgut, Sharp thresholds of graph properties, and the k-sat problem, *J. Amer. Math. Soc.* **12** (1999), 1017–1054. MR 1678031. Zbl 0932.05084. http://dx.doi.org/10.1090/S0894-0347-99-00305-7.

[9] A. Hildebrand and G. Tenenbaum, On integers free of large prime factors, *Trans. Amer. Math. Soc.* **296** (1986), 265–290. MR 0837811. Zbl 0601.10028. http://dx.doi.org/10.2307/2000573.

[10] C. Pomerance, The quadratic sieve factoring algorithm, in *Advances in Cryptology* (Paris, 1984), *Lecture Notes in Comput. Sci.* **209**, Springer-Verlag, New York, 1985, pp. 169–182. MR 0825590. Zbl 0596.10006. http://dx.doi.org/10.1007/3-540-39757-4_17.

[11] ———, The role of smooth numbers in number-theoretic algorithms, in *Proceedings of the International Congress of Mathematicians*, *Vol.* 1, 2 (Zürich, 1994), Birkhäuser, Basel, 1995, pp. 411–422. MR 1403941. Zbl 0854.11047.

[12] ———, Multiplicative independence for random integers, in *Analytic Number Theory*, *Vol.* 2 (Allerton Park, IL, 1995), *Progr. Math.* **139**, Birkhäuser, Boston, MA, 1996, pp. 703–711. MR 1409387. Zbl 0865.11085.

[13] ———, Smooth numbers and the quadratic sieve, in *Algorithmic Number Theory*: *Lattices*, *Number Fields*, *Curves and Cryptography*, *Math. Sci. Res. Inst. Publ.* **44**, Cambridge Univ. Press, Cambridge, 2008, pp. 69–81. MR 2467543. Zbl 1188.11065.

[14] R. D. Silverman, The multiple polynomial quadratic sieve, *Math. Comp.* **48** (1987), 329–339. MR 0866119. Zbl 0608.10004. http://dx.doi.org/10.2307/2007894.

[15] G. Tenenbaum, *Introduction to Analytic and Probabilistic Number Theory*, *Cambridge Stud. Adv. Math.* **46**, Cambridge Univ. Press, Cambridge, 1995. MR 1342300. Zbl 0880.11001.

(Received: August 9, 2010)
(Revised: February 7, 2011)

Georgia Institute of Technology, Atlanta, GA
E-mail: ecroot@math.gatech.edu

Université de Montréal, Montréal, Canada
E-mail: andrew@dms.umontreal.ca

University of Pennsylvania, Philadelphia, PA
E-mail: pemantle@math.upenn.edu

Georgia Institute of Technology, Atlanta, GA
E-mail: tetali@math.gatech.edu