Sharp Transitions in Making Squares

Ernie Croot[1]
School of Mathematics
Georgia Tech
Atlanta, GA 30332-0160
ecroot@math.gatech.edu

Andrew Granville [2]
Départment de mathématiques et de statistique
Université de Montréal
Montréal QC H3C 3J7, Canada
andrew@dms.umontreal.ca

Robin Pemantle[3]
Department of Mathematics
University of Pennsylvania
209 S. 33rd Street
Philadelphia, Pennsylvania 19104, USA
pemantle@math.upenn.edu

Prasad Tetali[4]
School of Mathematics and School of Computer Science
Georgia Tech
Atlanta, GA 30332-0160
tetali@math.gatech.edu

Abstract

In many integer factoring algorithms, one produces a sequence of integers (created in a pseudo-random way), and wishes to rapidly determine a subsequence whose product is a square (which we call a *square product*). In his lecture at the 1994 International Congress of Mathematicians, Pomerance observed that the following problem encapsulates all of the key issues: *Select integers $a_1, a_2, \ldots$, at random from the interval $[1, x]$, until some (non-empty) subsequence has product equal to a square. Find good estimate for the expected stopping time of this process.* A good solution to this problem should help one to determine the optimal choice of parameters for one's factoring algorithm, and therefore this is a central question.

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) $J_0 = J_0(x)$. Herein we determine this expected stopping time up to a constant factor, tightening 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. We will also confirm the well established belief that, typically, none of the integers in the square product have large prime factors.

We believe that there should, in fact, be a sharp threshold for this stopping time, that it should occur with probability $1 - o(1)$ after at least $\{e^{-\gamma} - o(1)\}J_0$ integers have been selected, but no more than $\{e^{-\gamma} + o(1)\}J_0$, with $J_0(x)$ as before.

Our proofs use methods somewhat different from previous articles on this subject. The heart of the proof of the 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 [12], the multiple polynomial quadratic sieve [17], and the number field sieve [2] – see [16] for a nice expository article on factoring algorithms) work by generating a pseudorandom sequence of integers $a_1, a_2, ...$, with each

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

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

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

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 non-trivial factor of n. If so, we have factored n.

In his lecture at the 1994 International Congress of Mathematicians, Pomerance [14, 15] observed that in the (heuristic) analysis of such factoring algorithms one assumes that the pseudo-random sequence $a_1, a_2, ...$ 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 $a_1, a_2, ... \leq x$ independently at random (that is, $a_j = m$ with probability $1/x$ for each integer m, $1 \leq m \leq x$), 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 $\leq y$, 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

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

In Pomerance's problem, let T be the smallest integer t for which $a_1, ..., a_t$ has a square dependence (note that T is itself a random variable). In 1985, Schroeppel gave a simple argument to justify that for any $\epsilon > 0$ we have

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

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

$$\mathrm{Prob}(T > J_0(x)^{1-\epsilon}) = 1 - o(1).$$

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 $T = J_0(x)^{1+o(1)}$. Pomerance asked in [3] whether there is a sharper transition, and we conjecture that T has a *sharp threshold*: This would mean that there exists a function $f(x)$ such that for every $\epsilon > 0$,

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

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

Conjecture 1.1 *For every $\epsilon > 0$ we have*

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

as $x \to \infty$, where $\gamma = 0.577...$ 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*

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

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

To obtain the lower bound in our theorem, we obtain a good upper bound on the expected number of sub-products of the large prime factors of the a_i's that equal a square, which allows us to bound the probability that such a sub-product exists, for $T < (\pi/4)(e^{-\gamma} - o(1))J_0(x)$. This is the "first moment method".

Schroeppel established his upper bound, $T \leq (1+o(1))J_0(x)$, 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 which 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 section 3.1 of [5] 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).

3

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 non-trivial 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 $a_1, ..., a_T$, including a_T, whose product equals a square.[5] It is not hard to show that this square product is T^2-smooth (see Section 3.2 of [5]); here we give a more precise idea of what I looks like:

Theorem 1.3

a) In the special case that for $\epsilon > 0$, conditional on the event $\{T < (\pi/4)(e^{-\gamma} - \epsilon)J_0(x)\}$, 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

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

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

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

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 $x \to \infty$.

---

[5]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.

There are now several theorems along the lines of Conjecture 1 in the literature, including some quite general approaches. Friedgut's theorem [8], characterizing a *coarse threshold* for monotone or symmetric[6] 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. Anyway, see [5] for further discussion of these issues.

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. Finally, in section 4, we develop our analysis of multiprime variations.

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 [10], which will come in useful later.

2.1 Classical smooth number estimates

From [10] we have that the estimate

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

---

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

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

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

and so

$$\Psi(x,y) \;=\; x\exp(-(u+o(u))\log u). \tag{7}$$

Now let

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

Then, using (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

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

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

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

6

By [10, Theorem 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

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

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 [10, Theorem 3] and (14) above, we have

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

Proposition 2.2 *Throughout the range (6), for any $1 \leq d \leq x$, we have*

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

where α is the solution to (11). In fact,

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

provided that

$$\frac{\log d}{\log u \log y + \sqrt{u \log u \log y}} \to \infty.$$

Proof. By (5), for $d = y^r$ with $0 \leq r \leq u/2$, we have

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

7

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 [18] gives that

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

so that the second term equals

$$-\int_{0}^{r}(\xi(u) - \xi(u-t))dt + O(r\log u/u).$$

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

$$\int_{0}^{r}(\xi(u) - \xi(u-t))dt \;=\; \int_{0}^{r}(r-v)\xi'(u-v)dv = \left(1 + O\left(\frac{1}{\log u}\right)\right)\int_{0}^{r}\frac{(r-v)}{(u-v)}dv$$

$$=\; \left(1 + O\left(\frac{1}{\log u}\right)\right)(r - (r-u)\log(1 - r/u)).\tag{19}$$

Combining this with the above yields that

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

$$+ O\left(\frac{r\log u}{u} + \frac{\log(u+1)}{\log y}\right)$$

$$=\; -\frac{r^2}{2u}\left\{1 + O\left(\frac{r}{u} + \frac{1}{\log u} + \frac{\log u}{r}\right)\right\} + O\left(\frac{\log(u+1)}{\log y}\right).$$

From (19) and the first equation here we find that this is negative provided $r \le u/2$ and $(\log u + \sqrt{u\log u/\log y})/r \to 0$, and is $o(1)$ in the complementary range.

If $d > \sqrt{x}$ we simply iterate the above result: The proposition follows by noting that $\alpha(x,y)$ is a decreasing function in x for fixed y, by definition. $\square$

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*

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

for all y outside of the range

$$y_0 \exp(-(1+\epsilon)\sqrt{\log y_0 \log\log y_0}) \leq y \leq y_0 \exp((1+\epsilon)\sqrt{\log y_0 \log\log y_0}); \qquad (20)$$

and

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

for all y outside of the range

$$y_0 \exp(-(c_3+\epsilon)\sqrt{\log y_0}) \leq y \leq y_0 \exp((c_3+\epsilon)\sqrt{\log y_0}). \qquad (21)$$

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 $g(u_1) \geq g(u_0)$ by definition of u_1; and $g(u_0) \geq g(u_1) + O(1/u_0)$ 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 $g'(v) = \rho'(v)/\rho(v) + v^{-2}\log x = -\xi(v) + v^{-2}\log x + O(\log v/v)$; so that, 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

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

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

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

and

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

which are the desired results. $\qquad \square$

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

Lemma 2.4 *We have*

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

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

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 $d = p_1 p_2 \ldots p_k$, where each p_j is a prime in $(y, My]$ 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 $k \geq 1$ and $\log M = o((\log x / \log\log x)^{1/4})$, as $x \to \infty$. Also

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

uniformly for $k \geq 1$ and $\log M = o(\log x / \log\log x)^{1/2}$, as $x \to \infty$.

Proof. We use (17) at most $2k$ times to obtain

$$\frac{\psi(x/(p_1 \cdots p_k), y_0)}{\psi(x, y_0)} = \frac{1}{(p_1 \cdots p_k)^\alpha} \left\{ 1 + O\left(\frac{k}{u_0} + \frac{k \log y_0}{y_0} \right) \right\} \sim \frac{(p_1 \cdots p_k)^{1-\beta}}{p_1 \cdots p_k}$$

where $\alpha(x, y_0) \geq \beta \geq \alpha(x/(p_1 \cdots p_k), y_0)$. If $u' = \log(x/(p_1 \cdots p_k) / \log(y_0)$ then $u' = u + O(k)$ and so $y_0^{1-\beta} = u_0 \xi(u_0)\{1 + O(k/u_0)\} = \log y_0 \{1 + O(k/u_0)\}$, by (15) and then Lemma 2.4. Hence we obtain (23) as $k^2 = o(u_0)$ and, in our range,

$$M^{k(1-\alpha)} = \exp(O(k \log M (\log\log y_0)/(\log y_0))) = 1 + o(1).$$

To obtain (24) we can use the same estimates but now we simply need $k/u_0 \to 0$ so that $y_0^{1-\beta} \leq (4/3) \log y_0$, and $\log M / u_0$ so that $M^{1-\beta} \leq (4/3)$. $\qquad\square$

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

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

where the sum is over primes q, and also that

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

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

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

and the result follows.

Lemma 2.6 *Let*

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

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

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

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

3 The lower bound for T in Theorem 1.2, and Theorem 1.3

3.1 Proof strategy

To establish that

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

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

3.2 Structure of a square product

We begin with the following proposition.

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

$$2 \; \leq \; |I| \; \leq \; \frac{\log x}{2 \log \log x} \quad \text{for which} \; \prod_{a \in I} a \; \text{is a square}$$

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

Proof. Suppose that $b_1, \ldots, b_k$ were chosen at random from $[1, x]$. The probability that $b_1 b_2 \ldots b_k$ is a square equals

$$x^{-k} |\{ b_1, \ldots, b_k \; \leq \; x \; : \; b_1 b_2 \ldots b_k \in \mathbb{Z}^2 \}|.$$

Now write each b_i uniquely as

$$b_i \; = \; c_i u_i^2, \quad \text{where } c_i \text{ is squarefree.}$$

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 $i, j = 1, \ldots, k$ and $i \neq j$, to be any satisfying the relations

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

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 $c_{i_1}, \ldots, c_{i_{2t}}$ (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:

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

Each $c_{i,j}$ is then the product of the primes dividing $c_1 \cdots c_k$ which 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 $c_1, \ldots, c_k$, the number of sequences $b_1, \ldots, b_k$ satisfying $b_i = c_i u_i^2$ is the number of possibilities for the numbers u_i, which is $\leq (x/c_i)^{1/2}$; and so, the probability that $b_1 \cdots b_k$ is a square is

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

$$\leq \; \frac{1}{x^{k/2}} \sum_{1 \leq i < j \leq k} \left(\sum_{c_{i,j} \leq x} \frac{1}{c_{i,j}} \right) \leq \frac{1}{x^{k/2}} (1 + \log x)^{k(k-1)/2} \tag{30}$$

12

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 $\prod_{i \in I} a_i \in \mathbb{Z}^2$ 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 $3 \leq k \leq \log x / 2 \log \log x$. $\qquad \square$

3.3 The main argument

In this subsection, we prove that

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

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 $1 \leq i \leq k$. 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

$$
\begin{aligned}
\text{Prob}(a_1 \ldots a_k \in \mathbb{Z}^2) \quad &\leq \quad \text{Prob}(d_1 \ldots d_k \in \mathbb{Z}^2) \\
&= \quad \sum_{\substack{d_1, \ldots, d_k \geq 1 \\ d_1 \ldots d_k \in \mathbb{Z}^2 \\ d_i = 1 \text{ or } p(d_i) > y}} \prod_{i=1}^{k} \frac{\Psi(x/d_i, y)}{x} \\
&\leq \quad \left(\{1 + o(1)\} \frac{\Psi(x, y)}{x} \right)^k \sum_{\substack{n=1 \text{ or } p(n) > y}} \frac{\tau_k(n^2)}{n^{2\alpha}}, \qquad (31)
\end{aligned}
$$

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 $\binom{J}{k} \leq (eJ/k)^k$; and so the expected number of k-tuples whose product is a square is

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

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

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

13

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

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

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

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

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

- Next, we consider the range

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

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

$$1 + \frac{\tau_k(p^2)}{p^{2\alpha}} + \frac{\tau_k(p^4)}{p^{4\alpha}} + \cdots \quad\sim\quad 1 + \frac{(k/p^\alpha)^2}{2!} + \frac{(k/p^\alpha)^4}{4!} + \cdots$$

$$= \frac{e^{k/p^\alpha} + e^{-k/p^\alpha}}{2}.$$

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

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

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

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

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

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

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).

14

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 $C = C(\epsilon)$ is sufficiently large then $e^{g(1,C)} < 4e^\gamma/\pi + \epsilon$ by Lemma 2.6 and, since y_0 maximizes $\Psi(x,y)/y$ for $y = y_0$, we deduce that (32) is at most

$$\leq ((1+\epsilon)4\eta e^\gamma/\pi)^k.$$

Therefore, if $\eta < (1-\epsilon)e^{-\gamma}\pi/4$, then this is less than $1/x^2$. $\qquad\square$

3.4 Proof of Theorem 1.3, part (a)

This last proof yields further useful information: If either $J < (\pi/4)(e^{-\gamma} - \epsilon)J_0(x)$, 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

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

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

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

By Lemma 2.3, this is at most

$$\left((4e^\gamma/\pi + \epsilon)(2/e^2 + o(1))\eta \right)^k < 1/2^k,$$

for sufficiently small $\epsilon > 0$, 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. $\qquad\square$

3.6 Proof of Theorem 1.3, part (c)

In the previous subsection we proved that

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

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 $d_{i_1} \ldots d_{i_l}$ is a subproduct which equals a square n^2, but such that no subproduct of this is a square. Then, with probability $1 - o(1)$, 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 $d_{i_1} \ldots d_{i_l}$ by $d_1 \ldots d_l$. 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 $n^2 = d_1 \ldots d_l > y^{2(l-1)}$: Create a graph G on the l vertices $v_1, \ldots, v_l$ where, for each prime p^q which (exactly) divides n, draw a total of q edges, placing an edge between pairs of vertices v_j for which p divides d_j. Now G is connected, since our square product is minimal, and so must have $\ge l - 1$ edges.

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 $d_1 \ldots d_l \ge y^{2l}\phi(x)^2$, 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 $(1 + \tau_l(p^2)/p^{2(\alpha-\theta)} + \tau_l(p^4)/p^{4(\alpha-\theta)} + \ldots)$.

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

$$
\tau_\ell(p^{2k}) = \binom{2k + \ell - 1}{2k} \le \begin{cases} \ell(\ell+1)/2, & \text{if } k = 1; \\ \frac{(2k+\ell-1)^{2k}}{(2k)!}, & \text{if } k \ge 2. \end{cases}
$$

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:

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

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

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

(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 $\theta \le 1/2$ by definition, so the above calculation becomes

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

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

$$= \frac{(\ell+1)\log \ell}{\log\left(\ell \log^2 y / 2 \log \ell\right)} = \ell \left(1 + O\left(\frac{\log \log \ell}{\log \ell}\right)\right) = \{1 + o_{\ell \to \infty}(1)\}\ell .$$

So putting (32) to use as explained above, the expected number of such l-tuples is

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

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

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

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 $\ll y/(\log y)^{\epsilon l/5}$. This is $o(1)$ provided $\ell \geq 6 \log y/(\epsilon \log \log y)$, 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 $d_1 \ldots d_l \geq y^{2l}$; hence $d_1 \ldots d_l = n^2 < y^{2l}$ 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

$$\sum_{p > y} \left(\frac{\tau_l(p^4)}{p^{4\alpha}} + \frac{\tau_l(p^6)}{p^{6\alpha}} + \ldots\right) \ll \frac{l^4}{(y/\log y)^3} \ll \frac{(\log y)^7}{y^3}.$$

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

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \square$

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 $a_1, a_2, \ldots$, chosen uniformly at random from $\{1, 2, \ldots, x\}$, 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 *k large primes variation*. We consider only rows in which at most k ones remain. 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, in addition to restricting k, we must restrict the columns: specifically, fixing $M > 0$, 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 (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 $\gamma_{m,M,k}$. For fixed m, M, k, these are ugly, but as $m, M, k \to \infty$, 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 $\{\gamma_{m,M,k}\}$. A result (Theorem 4.1) is then formulated in terms of these functions which 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 $\theta_{m,k}^{M,\eta}(\rho)$, 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, M, k 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, k and M.

$$
\begin{aligned}
\exp_k(z) &:= \sum_{j=0}^{k-1} \frac{z^j}{j!} \, ; \\
A_M(z) &:= \int_{1/M}^{1} \frac{1 - e^{-zt}}{t} \, dt \, .
\end{aligned}
$$

Clearly, as $k, M \to \infty$, we have the limits

$$\exp_k(z) \quad \uparrow \quad \exp(z);$$
$$A_M(z) \quad \uparrow \quad A(z) := \int_0^1 \frac{1 - e^{-zt}}{t} \, dt \, .$$

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

$$\gamma_{0,M,k}(u) \quad := \quad u \, ;$$
$$\gamma_{m+1,M,k}(u) \quad := \quad u \, \exp_k\left[A_M(\gamma_{m,M,k}(u))\right] \, . \tag{39}$$

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

$$\gamma_{M,k}(u) := u \, \exp_k\left[A_M(\gamma_{M,k}(u))\right] . \tag{40}$$

We now establish that $\gamma_{M,k}(u) < \infty$ except perhaps when $M = k = \infty$: we have $0 \le A_M(z) \le \log M$ for all z, so that $u < \gamma_{M,k}(u) \le Mu$ for all u; in particular $\gamma_{M,k}(u) < \infty$ if $M < \infty$. Also, $A(z) = \log z + O(1)$ which, along with (39), implies that $\gamma_{\infty,k}(u) \sim u(\log u)^{k-1}/(k-1)!$; in particular $\gamma_{\infty,k}(u) < \infty$. As $M, k \to \infty$, the fixed point $\gamma_{M,k}(u)$ increases to the fixed point $\gamma(u)$ of the map $z \mapsto u e^{A(z)}$, 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, k are such that*

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

then with probability approaching 1, as $x \to \infty$, among ηJ_0 uniform random samples from $\{1, \dots, x\}$, the y-smooth numbers up to My with at most k large primes 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 $\lim_{t \uparrow \eta_*} \int_0^t \frac{\gamma(u)}{u} \, du = 1$.

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 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 supp $(\mathcal{H}) \subseteq V$ (note: in the literature, often this language would imply 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* 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 $V_{q,1}, \ldots, V_{q,n(q)}$ 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 $j \leq n(q)$.*

A moment's thought shows that if $\mathcal{G}$ is a tree-like hypergraph rooted at p then the tree structure on 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 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 : $ supp $(\mathcal{G}) \to$ 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 (and 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 $\epsilon > 0$ such that there is a bijection $\phi : $ supp $(\mathcal{G}) \to$ supp $(\mathcal{H})$ inducing an isomorphism on the hypergraphs, and satisfying $|\phi(\rho) - \rho| \leq \epsilon$ for all $\rho \in$ 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,

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

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 $\epsilon > 0$ such that there is a coupling $(\mathcal{G}, \mathcal{G}')$ of μ and μ' for which the probability of $d(\mathcal{G}, \mathcal{G}') > \epsilon$ 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.

4.2 The random hypergraph $\mathcal{G}$ of (My)-smooth numbers

Before we get started, here are a few words on notation. As before, we are selecting random positive integers $\leq x$, with $y(x)$ 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 also choose an integer $k \geq 1$ and keep track only of numbers with at most k large prime factors (factors in the interval (y, My)); we may even choose $k = \infty$ in the range implied by the limitations given to the uniformity of (23). We will also specify an integer $m \geq 1$ which 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 which will depend on M, J, x, m, k and a large prime $p \in (y, My)$. The full notation for this will be $\mathcal{G}_{m,k,p}^{M,J,x}$. However, in most of the results and constructions that follow, k, 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 four parameters understood. In many of our lemmas, arises the phrase, "$f = o(1)$ 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 $f(M, J, x, m, k, p) < g(M_0, \eta, x, m, k)$ 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, k, 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 $p^i \mid X$ 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*

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

1. *For any finite set S of primes in (y, My) with $|S| \geq 2$, 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:

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

2. *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 $\mathcal{W}$, $\mathbb{P}_x(N(\mathcal{W}) \geq 2) \leq (\mathbb{E}_x N(\mathcal{W}))^2$.*

3. *For any $p \in (y, My)$, the probability that there will be a prime $q \neq p$ 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 $y < p, q \leq My$.*

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), 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 $\pi(y) \leq 2y / \log y$ instead of $\pi(y) \sim y / \log y$, gives (43).

The second statement follows because $N(\mathcal{W})$ has a binomial distribution. 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 $p_1 \neq p_2$. Let $\mathcal{W}_k$ denote the set of sets $S = \{p_1, p_2, \ldots, p_k\}$ of distinct primes between y and My and let $\mathcal{W} = \cup_{k \geq 2} \mathcal{W}_k$. By the second statement of this proposition, an upper bound for $H(p_1)$ 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 $S = \{p_1, p_2, \ldots, p_k\}$ to obtain

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

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

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

We sum this over all integers $k \geq 3$ so that

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

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

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

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

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 $X_1, \ldots, X_J$, and they will bear witness to the creation of pseudo-smooth numbers. They depend on the parameters M, J, x and k, which are fixed throughout the construction and suppressed in the notation. We remark that the definition makes sense for $k = \infty$.

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 $m \geq 1$ and $1 \leq j \leq J$ as follows.*

- *Let $T_0(p) := \{p\}$ and $\mathcal{G}_{0,p} := \emptyset$, taking $\mathrm{supp}\,(\mathcal{G}_{0,p}) = \{p\}$ by convention.*

- *For each $m \geq 1$, define $\mathcal{G}_{m,p}(0) := \mathcal{G}_{m-1,p}$. For $j \geq 1$, define $\mathcal{G}_{m,p}(j) := \mathcal{G}_{m,p}(j-1) \cup \{[X_j]\}$ if $[X_j]$ intersects $\text{supp}\,(\mathcal{G}_{m,p}(j-1))$ in a single element of $T_{m-1}(p)$ and $2 \leq |[X_j]| \leq k$. Otherwise, let $\mathcal{G}_{m,p}(j) := \mathcal{G}_{m,p}(j-1)$. Define $\mathcal{G}_{m,p} := \mathcal{G}_{m,p}(J)$. Define $T_m(p) := \text{supp}\,(\mathcal{G}_{m,p}) \setminus \text{supp}\,(\mathcal{G}_{m-1,p})$.*

Let U denote the set of primes q with $y < q < My$ such that $[X_j] = q$ for some $j \leq J$. Let $U_{m,p} := U \cap \text{supp}\,(\mathcal{G}_{m,p})$. 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 *For any η, M, x and p,*

$$\mathbb{E}_x\, |\mathcal{G}_{1,p}| \leq (2M-1)\eta\frac{y}{p}\,.$$

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 primes $p = p_1, p_2, \ldots, p_k$ in the range (y, My). The sum over ordered such sets is bounded above by the sum over ordered k-tuples in which repetition is allowed. Thus

$$\mathbb{E}_x|\mathcal{G}_{1,p}| \leq \sum_{k\geq 2} \frac{1}{(k-1)!} \sum_{p_2,\ldots,p_k} \mathbb{E}_x N(p, p_2, \ldots, p_k)$$

where the summand is zero, by convention, if there is a repetition. When there is no repetition, we obtain an estimate from (42), which implies the upper bound

$$\mathbb{E}_x|\mathcal{G}_{1,p}| \leq \sum_{k\geq 2} \frac{1}{(k-1)!} \sum_{p_2,\ldots,p_k} \eta\frac{y}{p} \prod_{i=2}^{k} \frac{\log y}{p_i}\,.$$

The inner sum factors as a power, yielding

$$\mathbb{E}_x|\mathcal{G}_{1,p}| \leq \eta\frac{y}{p} \sum_{k\geq 2} \frac{1}{(k-1)!} \left(\sum_{y<q<My} \frac{\log y}{q} \right)^{k-1}\,.$$

24

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

$$\mathbb{E}_x|\mathcal{G}_{1,p}| \leq \eta\frac{y}{p}\sum_{k\geq 2}\frac{(\log(2M))^{k-1}}{(k-1)!} = (2M-1)\eta\frac{y}{p}.$$

$\square$

Corollary 4.6

$$\mathbb{E}_x|\mathcal{G}_{1,p}|^2 \leq \mathbb{E}_x|\mathcal{G}_{1,p}| + (\mathbb{E}_x|\mathcal{G}_{1,p}|)^2$$

and

$$\mathbb{E}_x|\mathcal{G}_{m,p}| \leq (1+2\eta M)^m\frac{y}{p}.$$

Proof. For the first statement, note that for $S \neq T$, 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

$$\mathbb{E}_x|\mathcal{G}_{1,p}|^2 \leq \sum_{S,T}\mathbb{P}_x\left(S,T \in \mathcal{G}_{1,p}\right)$$

$$\leq \left(\mathbb{E}_x|\mathcal{G}_{1,p}|\right)^2 + \mathbb{E}_x|\mathcal{G}_{1,p}|.$$

For the second statment, 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 probaiblity $N(S)$. By induction, the mean number of such q is at most $(1+2\eta M)^{m-1}y/p$. Bounding the mean of each Poisson variable from above by $2\eta M$, we complete the induction. $\square$

The number of pseudo-smooths generated by time j, by definition, 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), $[X_j] = [X_i]$ for some $i < j$ (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 $\chi^{M,j}_{m,k,p}$ whose informal interpretation is that $\{p\}$ is in the span of $\langle[X_1],\ldots,[X_j]\rangle$ and that this fact is witnessed by classes $[X_i]$ of cardinality at most k, having indices $i \leq j$. A proposition immediately following the definition verifies the interpretation. The parameters k, 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)

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

2. 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 (i) $S \subseteq T_{r+1}(p) \cup \{q\}$ (that is, S is a hyperedge that appears first at distance $r + 1$ from p, and is a "child" of q), and (ii) the event $\chi(q')$ occurs for each $q' \in S$ other than q.

3. Finally, let $\chi(G, U)$ denote the event $\chi(G, U)(p)$.

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

Definition 4.9 (smooth 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,

$$\chi_{m,p} := \chi(\mathcal{G}_{m,p}, U_{m,p}, p)\,.$$

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 $\mathbf{V}$. In the following 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 $m \ge 1$, 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 $\ell(q) \ge 0$. If $\ell(q) = 0$ then $\chi_{m,p}(q)$ implies $[X_j] = \{q\}$ for some $j \le J$, which immediately implies $\{q\} \in \langle [X_1], \ldots, [X_j] \rangle$. Now suppose $\ell(q) \ge 1$. If $\chi_{m,p}(q)$ holds, let j satisfy (i) and (ii) of the definition with $q = p$. For each $q' \in [X_j]$ distinct from q, $\ell(q') \le \ell(q) - 1$, 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. $\square$

It follows from this that for any m, the number of linear dependences among $\{[X_1], \ldots, [X_J]\}$ is bounded from below by

$$\#\{j \le J : \text{ for all } p \in [X_j], \text{ the singleton } \{p\} \text{ is in the span } \langle [X_1], \ldots, [X_{j-1}] \rangle\}. \quad (44)$$

4.3 Construction of the limit object $\mathcal{H}_m$

An informal description of the limit object is as follows. 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:

(1) Countable additivity in S: if $\mathcal{A}$ is a collection of disjoint elements of $\mathcal{B}$ then $N\left(\bigcup_{S \in \mathcal{A}} S\right) = \sum_{S \in \mathcal{A}} N(S)$;

(2) Counting measure: $N(S)$ takes values in the nonnegative integers;

(3) Poisson distribution: for fixed S, the random variable $N(S)$ is distributed as a Poisson distribution with mean $\mu(S)$;

(4) 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 $\eta > 0$ and an integer $k \geq 2$. We construct a random hypergraph $\mathcal{H}_{m,p} = \mathcal{H}_{m,k,p}^{M,\eta}$ on a new probability space $(\Omega, \mathcal{F}, \mathbb{P})$ whose vertex set is the real interval $[1, M]$. The collection $[1, M]_j$ of subsets of $[1, M]$ of cardinality j may be identified with the sector $W_j \subseteq \mathbb{R}^j$ defined by

$$W_j := \{(\rho_1, \ldots, \rho_j) \in \mathbb{R}^j : 1 \leq \rho_1 < \cdots < \rho_j \leq M\} .$$

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

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

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 increasing limit $\mu^{+\rho}$ are supported on finite sets of 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 $\nu_\tau = \nu_{k,\tau}^{M,\eta}$ (as usual, we suppress quantities that are, for the moment, fixed) be the law of the points of a Poisson process with intensity $\mu^{+\rho}/\tau$. 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 non-atomicity 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 $\mathcal{H}_{m,\rho} = \mathcal{H}_{m,k,\rho}^{M,\eta}$, by recursion on m. For $m = 1$, choose $\mathcal{H}_{1,\rho}$ from the law ν_ρ. For $m \geq 1$, let $T_{m,\rho} = \text{supp}(\mathcal{H}_{m,\rho}) \setminus \text{supp}(\mathcal{H}_{m-1,\rho})$, taking $\text{supp}(\mathcal{H}_{0,\rho}) = \{\rho\}$ 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 $\tau \in \text{supp}(\mathcal{H}_{m,\rho})$ independently, with probability $1 - e^{-\eta/\tau}$.*

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

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

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 which will be used to compute the rate of accumulation of pseudo-smooth numbers.

Theorem 4.12 *Fix integers $m, k \geq 1$ and any real $M > 1$. Then*

$$\mathbb{P}_x(\chi_{m,k,p}^{M,j}) = (1 + o(1))\theta_{m,k}^{M,j/J_0}(p/y) \tag{45}$$

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

$$\mathbb{P}_x\left(\bigcap_{i=1}^r \chi_{m,k,p_i}^{M,j}\right) = (1 + o(1))\prod_{i=1}^r \theta_{m,k}^{M,j/J_0}(p_i/y), \tag{46}$$

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. In each case, the lemmas hold for any fixed k, and with $k = \infty$ in the range of uniformity given for (23).

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 (and even for $k = \infty$ in the range allowed by using (23)).

Next, let $\Xi = (\tau_1, \tau'_1] \times \cdots \times (\tau_n, \tau'_n]$ be any rectangular subset of the sector W_n and let Ξ_x denote the set of sets, S, of n 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

$$\mathbb{E}_x N(\sigma_p(\Xi_x)) \;=\; \sum_{S \in \sigma_p(\Xi_x)} \mathbb{E}_x N(S)$$

$$\sim\; \sum_{S \in \sigma_p(\Xi_x)} \eta \frac{y(\log y)^n}{p \prod_{q \in S} q} \, .$$

Factoring the sum of products gives the equivalent expression

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

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. We need to show that for any disjoint sets $\Xi^{(1)}, \ldots, \Xi^{(n)}$, 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 this 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,

$$\mathbb{P}_x(\mathcal{G}'_{1,p} \neq \mathcal{G}''_{1,p}) = O\left(\sqrt{J_0}\, \mathbb{P}_x(p \in [X_1])\right) = O\left(J_0^{-1/2}\, \mathbb{E}_x |\mathcal{G}'_{1,p}|\right) = O\left(J_0^{-1/2}\right), \quad (47)$$

by Corollary 4.6. $\qquad\square$

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

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

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 $p_1, \ldots, p_n$. All else is the same, once one observes that Poissonization gives (47) simultaneously for all $p_1, \ldots, p_n$. $\qquad\square$

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 $y^{-1}(\mathcal{G}_{m,p_i}^{M,j,x})_{1 \leq i \leq n}$ 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 $G_1, \ldots, G_r$ denote the subtrees of $\mathbf{T}_p(\mathcal{G}_{m,p})$ from the vertices $q_1, \ldots, q_r$ of $T_1(p)$. Let $\mathcal{G}(1), \ldots, \mathcal{G}(r)$ 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 $y^{-1}(\mathcal{G}(1), \ldots, \mathcal{G}(r))$ 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

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

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

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

30

By Corollary 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 $H_1, \ldots, H_r$) 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. $\qquad\square$

Lemma 4.16 *As $x \to \infty$, the distance in the weak metric between the random marked hypergraph $y^{-1}(\mathcal{G}^{M,j,x}_{m,p}, U_{m,p})$ and the random marked hypergraph $(\mathcal{H}^{M,j/J_0}_{m,p/y}, \tilde{U}_{m,p/y})$ 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*

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

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

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 $\mathrm{supp}\,(\mathcal{G}_{m,p})$. This is true since, in the limit $(x, y \to \infty$ and $J = \eta x \pi(y)/\psi(x,y))$, the events $|\{j : [X_j] = \{q_i\}, j = 1, \ldots, J\}|$ for fixed $q_1, q_2, \ldots, q_r$ 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 $\mathrm{supp}\,(\mathcal{G}_{m,p})$ are independent with probability $1 - e^{-\eta y/q}$, then the first part of the lemma is proven; the second part is analogous.

$\qquad\square$

Proof of Theorem 4.12. Begin with (45). 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 (45) 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 (45). The proof of (46) is identical, using the n-tuple convergence in Lemma 4.16 in place of convergence of the single marked hypergraph. $\qquad\square$

4.5 Computation of θ

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

Lemma 4.17
$$\theta_{m,k}^{M,\eta}(\rho) = 1 - e^{-\gamma_{m,M,k}(\eta)/\rho} \, .$$

Proof. The quantities M, η and k will be fixed throughout the proof, so we write θ_m for $\theta_{m,k}^{M,\eta}$. 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 $m \geq 1$. 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 $\tau \in S \setminus \{\rho\}$ that is not in $\tilde{U}_{m,\rho}$. We have, by induction,

$$
\begin{aligned}
1 - \theta_{m+1}(\rho) &= e^{-\eta/\rho} \, \mathbb{E}\left[\prod_{S \in \mathcal{H}_{1,\rho}} \left(1 - \prod_{\tau \in S \setminus \{\rho\}} \theta_m(\tau) \right) \right] \\
&= e^{-\eta/\rho} \, \mathbb{E}\left[\prod_{S \in \mathcal{H}_{1,\rho}} \left(1 - \prod_{\tau \in S \setminus \{\rho\}} \theta_m(\tau) \right) \right], \quad (48)
\end{aligned}
$$

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

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

Applying this to (48) 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 $j \geq 2$, the law of $S \setminus \{\rho\}$ on $\{|S| = j\}$ is $\eta \mu_{j-1}/\rho$. We may incorporate $-\eta/\rho$ as the $j = 1$ 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

$$\log(1 - \theta_{m+1}(\rho)) = -\frac{\eta}{\rho} \sum_{j'=0}^{k-1} \int \prod_{\tau \in S'} \theta_m(\tau) \, d\mu_{j'}(S) .$$

Here the primes are introduced to clarify the changes of variable $j' = j - 1, S' = S \setminus \{\rho\}$. We now drop the primes and observe that μ_j is $1/(j!)$ times a product measure. Therefore the integral of the product factors, yielding

$$
\begin{aligned}
\log(1 - \theta_{m+1}(\rho)) &= -\frac{\eta}{\rho} \sum_{j=0}^{k-1} \frac{1}{j!} \left(\int_1^M \theta_m(\tau) \frac{d\tau}{\tau} \right)^j \\
&= -\frac{\eta}{\rho} \exp_k \left(\int_1^M \theta_m(\tau) \frac{d\tau}{\tau} \right) .
\end{aligned}
$$

32

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

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

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

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

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

Lemma 4.18 *Fix any $\eta > \eta_*$. Then*

$$\theta_{m,k}^{M,\eta}(\rho) \to 1,$$

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

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

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

which, in turn, leads inductively to

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

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

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

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

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

4.6 Proof of main theorems

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

$$\theta_{m,k}^{M,\eta_*+\epsilon}(\rho) > \frac{3}{4} \quad \text{for all} \quad 1 \leq \rho \leq L := \exp\left(\frac{3}{\epsilon}\right).$$

Take M to be larger if necessary so that we may assume $M \geq L$. 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

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

Now let Y be the number of j in the interval $I := [(\eta + \epsilon)J_0, (\eta + 2\epsilon)J_0]$ 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 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,

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

The outer sum has at least ϵJ_0 terms, hence

$$\mathbb{E}_x Y \geq \frac{1}{2}(\epsilon J_0)\frac{\pi(y)}{J_0}(3\epsilon^{-1}) = \frac{3}{2}\pi(y). \tag{49}$$

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

$$\mathrm{Cov}\,(Y_i, Y_j) = o\left(\frac{\pi(y)^2}{J_0^2}\right).$$

Using, this lemma,

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

Together with (49), this implies that $\mathbb{P}_x(Y > \pi(y)) \to 1$. Recall from (44) that this implies more than $\pi(y)$ linear dependences among the classes $[X_j]$ with $j \leq (\eta_* + 2\epsilon)J_0$. Since $\epsilon > 0$ was arbitrary, this completes the proof of the theorem, modulo the lemma. $\square$

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 $[(\eta_* + \epsilon)J_0, (\eta_* + 2\epsilon)J_0]$. We can get much more reasonable values of m, M and k 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, M and k satisfying the inequality in the hypothesis of the theorem. Let

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

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

$$\liminf_{x \to \infty} \frac{\mathbb{E}_x Z}{\pi(y)} > 1 \,, \tag{50}$$

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

To prove (50), break down $\mathbb{E}Z_j$ according to the value of $[X_j]$ and using independence of X_j from $\chi_{m,k,p}^{M,j-1}$. This gives

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

The final equality above used both equation (23) and the formula (46) of Theorem 4.12. Continuing, we use the identity $\psi(x, y)/x = \pi(y)/J_0$, factor out this term, and rewrite the summand as a product:

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

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

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

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

$$\mathbb{E}_x Z_j \quad \sim \quad \frac{\pi(y)}{J_0} \exp\left(\frac{1}{y} \sum_{y < p < My} \frac{\log y}{p/y} \theta_{m,k}^{M,j/J_0}(p/y)\right)$$

$$\sim \quad \frac{\pi(y)}{J_0} \exp\left(\int_1^M \frac{1}{t} \theta_{m,k}^{M,j/J_0}(t) \, dt\right),$$

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

$$\frac{\mathbb{E}_x Z}{\pi(y)} \quad \sim \quad \int_0^\eta \exp\left(\int_1^M \frac{1}{t} \theta_{m,k}^{M,u}(t) \, dt\right) \, du$$

$$= \quad \int_0^\eta \frac{\gamma_{m+1}(u)}{u} \, du.$$

By the hypothesized inequality, the right-hand side is greater than 1, which establishes (50) and completes the proof of the theorem. □

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

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

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

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

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

uniformly for $1 \le i < j \le J$. 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 $\mathcal{G}_{m,k,p}^{M,i-1,x}$ for $p \in [X_i]$ and $\mathcal{G}_{m,k,p}^{M,j-1,x}$ for $p \in [X_j]$ are disjoint and tree-like with probability going to 1, and asymptotically independent. The same is true of the marked hypergraphs, by Lemma 4.16. Since Z_i is a bounded function of $[X_i]$ and the marked hypergraphs $(\mathcal{G}_{m,k,p}^{M,i-1,x}, U_{m,k,p}^{M,i-1,x})$ 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

$$c \frac{\pi(y)}{\Psi(x,y)/x} \sim \frac{y^2}{\log y \log \log y} \tag{51}$$

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 (though always with $m = \infty$, 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

```
The value of η such that there are ∼ π(y)
pseudosmooths amongst the aⱼ with j ≤ ηπ(y)x/Ψ(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 (51) by ηc, and determine that the new running time is given by (51), after solving (51) 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

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

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.

References

[1] M. Abramowitz and I. Stegun, Handbook of mathematical functions, Dover Publications, New York 1965.

[2] J. Buhler, H. W. Lenstra Jr., and C. Pomerance, *Factoring integers with the number field sieve*, Lecture Notes in Math, 1554, Springer, Berlin, 1993.

[3] N. Calkin, *Dependent sets of constant weight binary vectors*, Combin. Probab. Comp. **6** (1997), 263-271.

[4] R. Crandall and C. Pomerance, *Prime numbers; A computational perspective*, Springer, New York (2005).

[5] E. Croot, A. Granville, R. Pemantle, and P. Tetali, *Running time predictions for factoring algorithms*, Algorithmic Number Theory, ANTS VIII, Banff, Springer LNCS **5011** (2008), 1-36.

[6] J. D. Dixon, *Asymptotically fast factorization of integers*, Math. Comp. **36** (1981), 255-260.

[7] Richard Durrett, *Probability: Theory and Examples*, 3rd Edition.

[8] E. Friedgut, *Sharp Thresholds of Graph Properties, and the k-SAT Problem*, J. Amer. Math. Soc. **12** (1999), 1017-1054.

[9] A. Granville and K. Soundararajan, *Large Character Sums*, J. Amer. Math. Soc. **14** (2001), 365-397.

[10] A. Hildebrand and G. Tenenbaum, *On integers free of large prime factors*, Trans. Amer. Math. Soc **296** (1986), 265–290.

[11] Pierre Leroux, *Enumerative Problems Inspired by Mayer's Theory of Cluster Integrals*, Electronic Journal of Combinatorics, paper R32, May 14, 2004.

[12] C. Pomerance, *The quadratic sieve factoring algorithm.* Advances in cryptology, Paris (1984), 169-182.

[13] C. Pomerance, *The number field sieve* in Mathematics of Computation 1943–1993: a half century of computational mathematics (W. Gautschi, ed.), Proc. Symp. Appl. Math. **48**, Amer. Math. Soc., Providence (1994), 465 - 480.

[14] C. Pomerance, *The role of smooth numbers in number theoretic algorithms*, Proc. International Cong. of Mathematicians (Zurich, 1994), Birhäuser **1** (1995), 411 - 422.

[15] C. Pomerance, *Multiplicative independence for random integers*, Analytic number theory: Proceedings of a conference in honor of Heini Halberstam (eds., B.C. Berndt et. al.), Birhäuser **2** (1996), 703 - 711.

[16] C. Pomerance, *Smooth numbers and the quadratic sieve*, Proc. of an MSRI workshop, J. Buhler and P. Stevenhagen, eds. (to appear).

[17] R. Silverman, *The multiple polynomial quadratic sieve*, Math. Comp. bf 48 (1987), 329-339.

[18] G. Tenenbaum, Introduction to the analytic and probabilistic theory of numbers, Cambridge Univ. Press 1995.