A NON-UNIFORM BIRTHDAY PROBLEM WITH APPLICATIONS TO DISCRETE LOGARITHMS

STEVEN D. GALBRAITH AND MARK HOLMES

Abstract. We consider a generalisation of the birthday problem that arises in the analysis of algorithms for certain variants of the discrete logarithm problem in groups. More precisely, we consider sampling coloured balls and placing them in urns, such that the distribution of assigning balls to urns depends on the colour of the ball. We determine the expected number of trials until two balls of different colours are placed in the same urn. As an aside we present an amusing "paradox" about birthdays.

Keywords: birthday paradox, discrete logarithm problem (DLP), probabilistic analysis of randomised algorithms

1. Introduction

In the classical birthday problem one samples uniformly with replacement from a set of size N until the same value is sampled twice. It is known that the expected time at which this match first occurs grows as $\sqrt{\pi N/2}$. The word "birthday" arises from a common application of this result: the expected value of the minimum number of people in a room before two of them have the same birthday is approximately 23.94 (assuming birthdays are uniformly distributed over the year). The birthday problem can be generalised in a number of ways. For example, the assumption that births are uniformly distributed over the days in the year is often false. Hence, researchers have studied the expected time until a match occurs for general distributions. One can also generalise the problem to multi-collisions (e.g., 3 people having the same birthday) or coincidences among individuals of different "types" (e.g., in a room with equal numbers of boys and girls, when can one expect a boy and girl to share the same birthday). A good modern survey of results is DasGupta [6].

The main topic of this paper is matches of different types. This problem has important applications to computing discrete logarithms and our result has been used to obtain the results in [9, 10]. The problem will be stated in terms of sampling coloured balls and placing them in urns: What is the expected number of trials before there is an urn containing two balls of different colours? (We call such an event a *collision*.) Such problems have been considered by Nishimura and Sibuya [13, 14] and Selivanov [17]. If there are N urns, two colours, and balls are coloured independently with probability 0.5 and placed in urns uniformly with probability $1/N$ then the expected number of trials to get a collision is order $\sqrt{\pi N}$. An application is: if one has n boys and n girls in a room, what is the expected value of n before a girl and a boy share a birthday? The answer is 16.93 (the expected number of people in the room is $2n = 33.86$).

Described in terms of urn problems, the purpose of this paper is to consider collisions between balls of different colours when the balls are not assigned uniformly to urns. Some results of this type were obtained by Selivanov [17], however that paper assumes the distribution of assigning balls to urns is independent of the colour. For the applications to discrete logarithms it is necessary to consider the problem where the distribution of assigning balls to urns depends on the colour. In Section 6 we give a corollary of our result which is counter-intuitive and may be of independent interest.

It is well-known to probability theorists that such problems can be solved using the Stein-Chen method, and the techniques are quite standard. Nevertheless, the results are not immediate and it is necessary to check a number of technical details. For example, one must formulate the problem under consideration in the appropriate way to apply the Stein-Chen method, and one must ensure that the error terms are of lower order than the leading term etc. An alternative approach to these problems, that we did not investigate, is given by Flajolet, Gardy and Thimonier [7].

1.1. **The discrete logarithm problem.** The discrete logarithm problem (DLP) in a finite group G is: Given $g, h \in G$ to find an integer a, if it exists, such that $h = g^a$. This is a fundamental computational problem with applications to public key cryptography. There are a number of variants of the discrete logarithm problem. One example is the DLP in an interval: Given $g, h \in G$ and an integer N to find an integer a, if it exists, such that $h = g^a$ and $0 \leq a < N$. These variants arise in certain cryptosystems.

The best algorithms to solve the discrete logarithm problem in a general group originate in the work of Pollard [15, 16]. These algorithms exploit pseudorandom walks, are easily parallelised, and do not require large amounts of storage. Despite having exponential complexity, these algorithms are the best algorithms known to solve the discrete logarithm problem on general elliptic curves over finite fields (it is important to note that more efficient algorithms exist for certain other groups, such as large subgroups of the multiplicative group of a finite field). Recent work extending and improving Pollard's algorithms for certain variants of the discrete logarithm problem (such as the DLP in an interval) has been performed by the first author and his collaborators [8, 9, 10]. The analysis of these algorithms requires the generalisation of the birthday problem mentioned above. The purpose of this paper is to state and prove a theorem which can be used to analyse all algorithms of this type. Our result allows the expected running time of the algorithms in [9, 10] to be determined (to leading order) and hence one can determine optimal choices of parameters for these algorithms.

1.2. **Statement of the main result.** We now present our main result and the assumptions which are required to formulate the problem. The assumptions are supposed to capture any scenario that will arise in the analysis of algorithms for the discrete logarithm problem (not just the cases in [9, 10]).

Assumption 1 (Colour selection)**.** *We assume that there are $C \in \mathbb{N}$ different colours of balls. The k-th ball sampled has probability $\mathsf{r}_{k,c}$ of being colour c (independent of all previous selections) where, for every $c = 1, 2, \ldots, C$, $\mathsf{q}_c := \lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} \mathsf{r}_{k,c}$ exists, and $\mathsf{q}_1 \geq \mathsf{q}_2 \geq \cdots \geq \mathsf{q}_C > 0$. Let $b_{n,c} = \mathsf{q}_c - n^{-1} \sum_{k=1}^{n} \mathsf{r}_{k,c}$. Note that $b_{n,c}$ denotes the deviation of the average probability of type c balls up to time n from the asymptotic average. We assume that there is a constant K such that $|b_{n,c}| \leq K/n$ for all c.*

Assumption 2 (Urn selection)**.** *There are $N \in \mathbb{N}$ distinct urns (our results will be asymptotic as $N \to \infty$). If the k-th ball has colour c then the probability that it is put in urn a is $\mathsf{q}_{c,a}(N)$ (i.e., independent of previous colour and urn selections and of k). There exists $d > 0$ such that for every $1 \leq c \leq C$ and $1 \leq a \leq N$,*

$$0 \leq \mathsf{q}_{c,a} \leq d/N.$$

There exist constants $\alpha, \mu > 0$ such that the set $S_N := \{1 \leq a \leq N : \mathsf{q}_{1,a}, \mathsf{q}_{2,a} \geq \mu/N\}$ is such that $|S_N| \geq \alpha N$. This says that there is a set $S_N \subset \{1, \ldots, N\}$ that is "large" (i.e., $|S_N| \geq \alpha N$) and such that, for all subsets $S \subseteq S_N$, if a ball of colour one or two is chosen then the probability that the ball is put in S is at least proportional to the size of S (essentially this says that balls of different colour do not go into disjoint sets of urns).

Define

$$A_N = \sum_{c=1}^{C} \mathsf{q}_c \left(\sum_{\substack{c'=1,\\ c' \neq c}}^{C} \mathsf{q}_{c'} \left(\sum_{a=1}^{N} \mathsf{q}_{c,a} \mathsf{q}_{c',a} \right) \right). \tag{1}$$

This is the limiting (as $n \to \infty$) probability that two balls (n and $n + n'$ for $n' > 0$) are given different colours but assigned to the same urn.

Observe that by the above assumptions

$$A_N \;\geq\; \mathsf{q}_1 \mathsf{q}_2 \sum_{a \in S_N} \mathsf{q}_{1,a} \mathsf{q}_{2,a} \;\geq\; \mathsf{q}_1 \mathsf{q}_2 \alpha N \left(\frac{\mu}{N} \right)^2$$

whence there exists a constant $f > 0$ (independent of N) such that $A_N \geq f/N$. Similarly

$$A_N \leq \left(\sum_{a=1}^{N} \frac{d^2}{N^2} \right) \sum_{c=1}^{C} \mathsf{q}_c \left(\sum_{\substack{c'=1,\\ c' \neq c}}^{C} \mathsf{q}_{c'} \right) \leq \frac{d^2}{N}.$$

This says that $A_N = \Theta(N^{-1})$, where the notation $\Theta(x)$ denotes some quantity satisfying $\alpha_1 x \le \Theta(x) \le \alpha_2 x$ for some constants $0 < \alpha_1 < \alpha_2$. For $x \ge 0$ we write $O(x)$ for some quantity satisfying $|O(x)| \le \beta x$ for some constant $\beta > 0$.

For notational convenience, we assume that the random ball and urn selections are defined for all N simultaneously on a common probability space with probability measure $\mathbb{P}$. For example, this can be achieved by letting $\{U_i\}_{i\in\mathbb{N}}$ and $\{V_i\}_{i\in\mathbb{N}}$ be independent standard uniform random variables under the probability measure $\mathbb{P}$ and choosing ball k to be of colour c and put in urn a if and only if

$$U_k \in \Big(\sum_{c'=1}^{c-1} \mathsf{r}_{k,c'}, \sum_{c'=1}^{c} \mathsf{r}_{k,c'} \Big] \quad \text{and} \quad V_i \in \Big(\sum_{a'=1}^{a-1} \mathsf{q}_{c,a'}(N), \sum_{a'=1}^{a} \mathsf{q}_{c,a'}(N) \Big].$$

Note that this is an event of probability $\mathsf{r}_{k,c}\mathsf{q}_{c,a}(N)$. In this paper $\mathbb{P}(X = x)$ denotes the probability that a random variable X takes a value $x \in \mathbb{R}$, and we write $\mathbb{E}[X]$ for the expected value of X with respect to $\mathbb{P}$. For events A and B we write $\mathbb{P}(A, B)$ for $\mathbb{P}(A \cap B)$.

We can now state our main result.

Theorem 1. *Let Z_N be the first time that there are two balls of different colours in the same urn, under Assumptions 1 and 2. Let A_N be as in equation (1). Then*

$$\mathbb{E}[Z_N] = \sqrt{\frac{\pi}{2A_N}} \; + \; O(N^{1/4})$$

as $N \to \infty$ and the implied constant in the O depends on $C, \mathsf{q}_c, d, K, \alpha, \mu$ but does not depend on N or the values of $\mathsf{q}_{c,a}$.

For the error term one can group the colours into two disjoint groups and call one of them "colour 1" and the other "colour 2". But to get the right leading term it is necessary to keep the colours separate.

2. Method of Proof

A powerful method employed many times [2, 3, 6] when studying variants of the birthday problem is to use Poisson approximation and the Stein-Chen method [5, 1]. The method gives bounds on the error involved in approximating the distribution of a sum of dependent Bernoulli random variables by the Poisson distribution. For our particular problem, the basic idea is that when N is large, collision probabilities are "close" to a binomial distribution for moderate numbers ($n \ll N$) of trials. The number of collisions after $n \ll N$ trials should therefore behave like a Poisson random variable, and we are asking for the typical number of trials required to get a collision.

To state the form of the error bounds, we closely follow Section 10 of Chatterjee, Diaconis and Meckes [4] (also see Arratia, Goldstein and Gordon [1]). Let I be a finite set, and suppose that $\{X_i\}_{i\in I}$ are random variables on a common probability space, taking values in $\{0, 1\}$. For $i, j \in I$ let

$$p_i = \mathbb{P}(X_i = 1), \quad \text{and} \quad p_{i,j} = \mathbb{P}(X_i = 1, X_j = 1). \tag{2}$$

For $i \in I$ let $N_i \subset I$ be such that X_i and $\{X_j : j \in I \setminus N_i\}$ are independent. Define

$$\mathrm{Err}(I) = \sum_{i\in I} \sum_{j\in N_i - \{i\}} p_{i,j} + \sum_{i\in I} \sum_{j\in N_i} p_i p_j.$$

Recall that a random variable Y has a Poisson distribution with parameter λ, (i.e., $Y \sim \mathrm{Pois}(\lambda)$) if

$$\mathbb{P}(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k \in \mathbb{Z}_{\ge 0}. \tag{3}$$

Theorem 15 of [4] is as follows.

Theorem 2. *Let $\{X_i\}_{i\in I}$ be random variables on a common probability space, taking values in $\{0, 1\}$, and let $\lambda = \sum_{i\in I} p_i$ and $W = \sum_{i\in I} X_i$. Then for $Y \sim Pois(\lambda)$,*

$$\tfrac{1}{2} \sum_{k\in\mathbb{Z}_{\ge 0}} \big|\mathbb{P}(W = k) - \mathbb{P}(Y = k)\big| \le \min\{1, 1/\lambda\} Err(I). \tag{4}$$

This says that the total variation (which is called the statistical difference in the cryptography community) of the distribution of W and the Poisson distribution with parameter λ is bounded above by $\min\{1, 1/\lambda\}\mathrm{Err}(I)$. As we will see in Section 4, our main result follows from Theorem 2 for a suitable choice of the random variables X_i. Most of our work is devoted to controlling the error terms.

3. Recovering a Theorem of Selivanov

As a sanity check, we show how a special case of the result of Selivanov [17] is also a special case of our Theorem 1. Selivanov considers balls coloured with colour c with probability q_c and placed in urn a with probability p_a. So, translating to our notation, $\mathsf{q}_c = q_c$, $b_{n,c} = 0$ and $\mathsf{q}_{c,a} = p_a$. Selivanov defines, for $k \in \mathbb{N}$, $v_k = \sum_{a=1}^{N} p_a^k$ and $w_k = \sum_{c=1}^{C} q_c^k$. Selivanov makes the assumption that $v_2 = o(1)$ (this is weaker than our assumption $\mathsf{q}_{c,a} \leq d/N$ which implies $v_2 \leq d^2/N$). Hence, to apply our result we must impose the stronger condition $p_a \leq d/N$. Since p_a does not depend on the colour, the existence of α and μ is easy to check. Theorem 4.1 of Selivanov [17] is that the expected value of the number of balls assigned to get a collision is

$$\sqrt{\frac{\pi}{2v_2(1-w_2)}}(1+o(1)).$$

In our case we find that

$$A_N = \left(\sum_{c=1}^{C} q_c \sum_{c'=1, c' \neq c}^{C} q_{c'}\right)\left(\sum_{a=1}^{N} p_a^2\right)$$

which is easily checked to be $(1-w_2)v_2$. Hence, we obtain Selivanov's result with a stronger error bound.

One of the contributions to the error term in Theorem 2 is the quantity $p_{i,j}$. Later in this paper the analogous quantity will be denoted $p_{N,(i,i'),(j,j')}$, and its value is bounded in the proof of Lemma 2. In Selivanov's notation this value is $v_3(1-2w_2+w_3)$, which appears in equation (3.6) of Section 3 of [17]. Poisson approximation is used in that section of [17].

4. Proof of Theorem 1

For $i \in \mathbb{N}$ write K_i for the colour of ball i and $U_{N,i}$ for the urn into which it has been put. Set $I(n) = \{(i,i') : 1 \leq i < i' \leq n\}$ so that $\#I(n) = \binom{n}{2}$. Define, for $(i,i') \in I(n)$,

$$X_N(i,i') = \begin{cases} 1 & \text{if } U_{N,i} = U_{N,i'} \text{ and } K_i \neq K_{i'} \\ 0 & \text{otherwise.} \end{cases}$$

In other words, the random variable $X_N(i,i')$ takes the value 1 if and only if the i-th and i'-th balls form a collision (it will always be understood that this is a rare event).

Define

$$W_N(n) = \sum_{(i,i') \in I(n)} X_N(i,i'), \quad \text{and} \quad Z_N = \inf\{n \in \mathbb{N} : W_N(n) > 0\}.$$

Note that $W_N(0) = 0$.

For $n \in \mathbb{Z}_{\geq 0}$, the following equalities between events hold: $\{Z_N = n+1\} = \{W_N(n+1) = 1\} \cap \{W_N(n) = 0\}$ (this is the event that there is no collision among the first n balls, but the $n+1$-th ball leads to a collision) and $\{Z_N > n\} = \{W_N(n) = 0\}$ (this is the event that there is no collision among the first n balls). Our problem is to determine $\mathbb{E}[Z_N]$ to leading order as $N \to \infty$. We use Theorem 2 to answer this, and our notation is intended to be analogous to the notation used in [4]. Recall that the expected value of Z_N is

$$\mathbb{E}[Z_N] = \sum_{n=0}^{\infty} n\mathbb{P}(Z_N = n) = \sum_{n=0}^{\infty} \mathbb{P}(Z_N > n) = \sum_{n=0}^{\infty} \mathbb{P}(W_N(n) = 0).$$

The proof determines an approximation to the sum on the right hand side.

The proof is broken down into a number steps. The first step is to determine the various probabilities and the error term in Theorem 2. We determine the expected value by splitting the sum $\sum_{n=0}^{\infty} \mathbb{P}(W_N(n) = 0)$ into two parts: first the sum of n from 1 to $M = N^{1/2+1/16}$ (which gives the leading term) and then the sum from $M+1$ to ∞ (which is the "tail"). A big part of the proof (see Section 4.2) is giving a bound for the tail.

For the convenience of the reader we give a table of the notation in Figure 1.

N	Number of urns
C	Number of colours of balls
$\mathsf{r}_{k,c}$	Probability the k-th ball sampled has colour c
$\mathsf{q}_{c,a}(N)$	Probability, given that the k-th ball has colour c, that it is put in urn a
d	Constant such that $0 \leq \mathsf{q}_{c,a} \leq d/N$
q_c	Asymptotic average value for $\mathsf{r}_{k,c}$
$b_{n,c}$	Deviation of the average probability of colour c balls up to time n from the asymptotic average
K	Constant such that $\lvert b_{n,c}\rvert \leq K/n$ for all c
α, μ, S_N	The set $S_N = \{1 \leq a \leq N : \mathsf{q}_{1,a}, \mathsf{q}_{2,a} \geq \mu/N\}$ satisfies $\lvert S_N\rvert \geq \alpha N$
A_N	Defined in equation (1)
$I(n)$	$\{(i,i') : 1 \leq i < i' \leq n\}$
$X_N(i,i')$	Random variable which is 1 if and only if the i-th and i'-th balls have different colours but are put in the same urn
$W_N(n)$	$\sum_{(i,i')\in I(n)} X_N(i,i')$
Z_N	$\inf\{n \in \mathbb{N} : W_N(n) > 0\}$
M	$N^{1/2+1/16}$
$p_{N,(i,i')}$	$\mathbb{P}(X_N(i,i') = 1)$
$N_{(i,i')}$	Set of indices $(j,j') \in I(n)$ such that $X_N(i,i')$ and $X_N(j,j')$ are not independent
$\lambda_N(n)$	Defined in Lemma 1
$p_{N,(i,i'),(j,j')}$	$\mathbb{P}(X_N(i,i') = 1$ and $X_N(j,j') = 1)$
$\mathrm{Err}_N(n)$	Defined in equation (6)
$V_{n,N}$, $V'_{n,N}$	Defined in Section 4.2
ϵ	Small quantity used in Section 4.2
δ	$N^{-1/2}$
$Q_N(\delta)$, $P_1(n,N,\delta)$, $P_2(n,N,\delta)$	Defined in Section 4.2

FIGURE 1. Table of notation

4.1. **Determining Probabilities and Error Terms.** We use notation analogous to that used in Theorem 2. In particular, the random variables $X_N(i,i')$ are defined for $(i,i') \in I(n)$ and we denote $\mathbb{P}(X_N(i,i') = 1)$ by $p_{N,(i,i')}$, and $\mathbb{P}(X_N(i,i') = 1, X_N(j,j') = 1)$ by $p_{N,(i,i'),(j,j')}$.

By Assumptions 1 and 2, using the independence of ball and urn choices we have

$$\begin{aligned}\mathbb{P}(X_N(i,i') = 1) &= \sum_{a=1}^{N}\sum_{c=1}^{C} \mathbb{P}(U_{N,i} = a, K_i = c)\mathbb{P}(U_{N,i'} = a, K_{i'} \neq c) \\ &= \sum_{a=1}^{N}\sum_{c=1}^{C} \mathsf{r}_{i,c}\mathsf{q}_{c,a} \sum_{\substack{c'=1,\\ c'\neq c}}^{C} \mathsf{r}_{i',c'}\mathsf{q}_{c',a}.\end{aligned}$$

So that

$$p_{N,(i,i')} = \mathbb{P}(X_N(i,i') = 1) = \sum_{a=1}^{N}\sum_{c=1}^{C}\sum_{\substack{c'=1,\\ c'\neq c}}^{C} \mathsf{r}_{i,c}\mathsf{q}_{c,a}\mathsf{r}_{i',c'}\mathsf{q}_{c',a}. \tag{5}$$

Lemma 1. *Define*

$$\lambda_N(n) = \sum_{(i,i')\in I(n)} p_{N,(i,i')}.$$

Then

$$\lambda_N(n) = \frac{n^2}{2}A_N + O(n/N)$$

where the term $O(n/N)$ is a real number which is positive or negative but whose absolute value is bounded by cn/N for some constant $c > 0$.

Proof. It follows from the definitions and (5) that

$$\begin{aligned}
\lambda_N(n) &= \sum_{(i,i')\in I(n)} \sum_{a=1}^{N}\sum_{c=1}^{C}\sum_{c'\neq c}^{C} \mathsf{q}_{c,a}\mathsf{q}_{c',a}\mathsf{r}_{i,c}\mathsf{r}_{i',c'} \\
&= \frac{1}{2}\sum_{a=1}^{N}\sum_{c=1}^{C} \mathsf{q}_{c,a}\sum_{c'\neq c}\mathsf{q}_{c',a}\left[n\mathsf{q}_c n\mathsf{q}_{c'} - \left(n\mathsf{q}_c n\mathsf{q}_{c'} - \sum_{i=1}^{n}\mathsf{r}_{i,c}\sum_{i'=1}^{n}\mathsf{r}_{i',c'}\right) - \sum_{i=1}^{n}\mathsf{r}_{i,c}\mathsf{r}_{i,c'}\right] \\
&= \frac{n^2}{2}A_N - \frac{1}{2}\sum_{a=1}^{N}\sum_{c=1}^{C} \mathsf{q}_{c,a}\sum_{c'\neq c}\mathsf{q}_{c',a}\left[nb_{n,c}n\mathsf{q}_{c'} + nb_{n,c'}\sum_{i=1}^{n}\mathsf{r}_{i,c} + \sum_{i=1}^{n}\mathsf{r}_{i,c}\mathsf{r}_{i,c'}\right] \\
&= \frac{n^2}{2}A_N - \frac{1}{2}\sum_{a=1}^{N}\sum_{c=1}^{C} \mathsf{q}_{c,a}\sum_{c'\neq c}\mathsf{q}_{c',a}\left[nb_{n,c}n\mathsf{q}_{c'} + nb_{n,c'}n\mathsf{q}_c - nb_{n,c'}nb_{n,c} + \sum_{i=1}^{n}\mathsf{r}_{i,c}\mathsf{r}_{i,c'}\right] .
\end{aligned}$$

We now want to bound the error term. Since $|b_{n,c}| \leq K/n$ and $\mathsf{r}_{i,c} \leq 1$, we find that

$$\left|nb_{n,c}n\mathsf{q}_{c'} + nb_{n,c'}n\mathsf{q}_c - nb_{n,c'}nb_{n,c} + \sum_{i=1}^{n}\mathsf{r}_{i,c}\mathsf{r}_{i,c'}\right| < n\mathsf{q}_{c'}K + n\mathsf{q}_c K + K^2 + n$$

which is $O(n)$. There are $NC(C-1)$ terms being added, and $\mathsf{q}_{c,a} < d/N$, hence the error term is $O(n/N)$. It follows that $\lambda_N(n) = \frac{n^2}{2}A_N + O(n/N)$. □

Poisson approximation already suggests the distribution to be like $e^{-\lambda_N(n)}$ and so the expected value of n should be like $\sqrt{\pi/(2A_N)}$. However, it is important to keep track of the error terms since n/N is bigger than the leading term when n is large.

Note that $X_N(i,i')$ and $X_N(j,j')$ are independent if and only if $\{i,i'\} \cap \{j,j'\} = \emptyset$. In other words, whether or not the i and i'-th balls are a collision is independent of whether or not the j and j'-th balls are a collision when $\{i,i'\} \cap \{j,j'\} = \emptyset$. Furthermore, it is clear in this case that, for any set $J \subseteq I(n)$, $X_N(i,i')$ is independent of $\{X_N(j,j') : (j,j') \in J\}$ if and only if $\{i,i'\} \cap (\cup_{(j,j')\in J}\{j,j'\}) = \emptyset$.

Hence, we have

$$N_{(i,i')} = \{(j,j') \in I(n) : i = j \text{ or } i = j' \text{ or } i' = j \text{ or } i' = j'\}$$

and $|N_{(i,i')}| \leq 4n$. Now we determine $p_{N,(i,i'),(j,j')}$, which is only non-trivial when $(j,j') \in N_{(i,i')}$ (in the other case, $X_N(i,i')$ and $X_N(j,j')$ are independent and $p_{N,(i,i'),(j,j')} = p_{N,(i,i')}p_{N,(j,j')}$).

Lemma 2. *Let notation be as above and $(i,i') \in I(n)$. Let $(j,j') \in N_{(i,i')}$. Then*

$$p_{N,(i,i'),(j,j')} = \mathbb{P}(X_N(i,i') = 1 \text{ and } X_N(j,j') = 1) \leq C(C-1)^2 d^3/N^2.$$

Proof. Without loss of generality suppose $j = i$. So $U_{N,i} = U_{N,i'} = U_{N,j} = U_{N,j'}$ (which are all equal to some value $1 \leq a \leq N$) and $K_i = K_j$. It follows that $K_{i'} \neq K_i$ and $K_{j'} \neq K_i$ (the values of $K_i, K_{i'}$ and $K_{j'}$ are called c, c' and c'' below). Therefore

$$p_{N,(i,i'),(j,j')} = \sum_{c=1}^{C}\mathsf{r}_{i,c}\sum_{\substack{c'=1,\\ c'\neq c}}^{C}\mathsf{r}_{i',c'}\sum_{\substack{c''=1,\\ c''\neq c}}^{C}\mathsf{r}_{j',c''}\sum_{a=1}^{N}\mathsf{q}_{c,a}\mathsf{q}_{c',a}\mathsf{q}_{c'',a}.$$

Since $0 \leq \mathsf{r}_{i,c} \leq 1$ and $0 \leq \mathsf{q}_{c,a} \leq d/N$ the result follows. □

Define, as in Theorem 2,

$$\text{Err}_N(n) = \sum_{(i,i')\in I(n)}\sum_{(j,j')\in N_{(i,i')}-\{(i,i')\}} p_{N,(i,i'),(j,j')} + \sum_{(i,i')\in I(n)}\sum_{(j,j')\in N_{(i,i')}} p_{N,(i,i')}p_{N,(j,j')}. \tag{6}$$

We set $\text{Err}_N(0) = \text{Err}_N(1) = 0$. We now give some bounds which will be used later.

Lemma 3. *Let notation be as above and conditions as in Theorem 1. Then*

$$p_{N,(i,i')} = O(1/N)\ , \quad p_{N,(i,i'),(j,j')} = O(1/N^2)\ , \quad A_N = \Theta(1/N) \quad \text{and} \quad Err_N(n) = O(n^3/N^2).$$

Proof. We have

$$p_{N,(i,i')} = \mathbb{P}(X_N(i,i') = 1) = \sum_{a=1}^{N}\sum_{c=1}^{C}\sum_{\substack{c'=1,\\ c'\neq c}}^{C} \mathsf{r}_{i,c}\mathsf{q}_{c,a}\mathsf{r}_{i',c'}\mathsf{q}_{c',a}.$$

Since $0 \leq \mathsf{r}_{i,c} \leq 1$ and $0 \leq \mathsf{q}_{c,a} < d/N$ this is at most $C(C-1)N(d/N)^2$ which is $O(1/N)$. The second statement is Lemma 2. The claim about A_N has already been established in Section 1. The final result follows from the first, and the facts that $|I(n)| = \binom{n}{2} < n^2/2$ and $|N_{(i,i')}| \leq 4n$. □

We have shown in Lemma 3 that $\mathrm{Err}_N(n) \leq c_1 n^3/N^2$ for some constant $c_1 > 0$ and for sufficiently large N. Taking $M = N^{1/2+1/16}$ gives

$$\sum_{n=0}^{M} \mathrm{Err}_N(n) \leq \tfrac{c_1}{N^2}\sum_{n=0}^{M} n^3 = \frac{c_1 M^2(M+1)^2}{4N^2} = O(N^{4/16}) = O(N^{1/4}). \tag{7}$$

This explains why we take $M = N^{1/2+1/16}$. One could try to handle slightly larger values using the bound $\min\{1, 1/\lambda_N(n)\}\mathrm{Err}_N(n)$, but we prefer not to do this.

Obtaining the leading term of Theorem 1 using Poisson approximation is given in Section 4.3. Before then we need to bound the contribution to the expected value coming from values $n > M$. This is the aim of the next section.

4.2. **Bounding the Tail.** The aim of this section (which is the most detailed part of the paper) is to show that $\sum_{n>M}\mathbb{P}(X_N > n) = O(1)$.

We need to bound $\mathbb{P}(Z_N > n)$ directly for $n \geq M$. Recall that there exists a subset S_N of the urns of size at least αN and such that $\mathsf{q}_{1,a}, \mathsf{q}_{2,a} \geq \mu/N$ for all $a \in S_N$. Given any such S_N, let $V_{n,N}$ be the random variable giving the number of urns in S_N containing a ball of colour 1 once n balls have been put in urns, and let $V'_{n,N} \leq V_{n,N}$ be the number of urns in S_N containing a ball of colour 1 that also contain a ball of some other colour. The dependence of these quantities on S_N will be left implicit. For $\delta > 0$ we define

$$P_1(n,N,\delta) = \mathbb{P}(V_{n,N} \leq \delta|S_N|), \quad \text{and} \quad P_2(n,N,\delta) = \mathbb{P}(V_{n,N} > \delta|S_N|, V'_{n,N} = 0).$$

The first event says that after n balls have been put in urns, the collection of urns in S_N containing a ball of colour 1 is a small subset of S_N. The second event says that this collection is not so small, yet there is still no collision of colour 1 with some other colour in S_N. Then

$$\mathbb{P}(Z_N > n) \leq P_1(n,N,\delta) + P_2(n,N,\delta).$$

Lemma 4. *For every $0 < \epsilon, \delta < 1$ such that $\epsilon < \min\{\mathsf{q}_1, \mathsf{q}_2\}$ and all n sufficiently large (depending on ϵ),*

$$P_1(n,N,\delta) \ \leq\ 2e^{-\frac{\epsilon^2}{2}n} + \sum_{S\in Q_N(1-\delta)}\left(1 - \sum_{a\in S}\mathsf{q}_{1,a}(N)\right)^{\lceil \mathsf{q}_1-\epsilon\rceil n},$$

and

$$P_2(n,N,\delta) \ \leq\ 2e^{-\frac{\epsilon^2}{2}n} + \max_{S\in Q_N(\delta)}\left(1 - \sum_{a\in S}\mathsf{q}_{2,a}(N)\right)^{\lceil \mathsf{q}_2-\epsilon\rceil n}.$$

Proof. The probability that the kth ball has colour 1 is $\mathsf{r}_{k,1}$. Let Y_n be the number of balls of colour 1 chosen by time n. Let $\epsilon > 0$ and choose n_0 such that for all $n \geq n_0$, $|n^{-1}\sum_{k=1}^{n}\mathsf{r}_{k,1} - \mathsf{q}_1| < \epsilon/2$. Then for all $n \geq n_0$,

$$\mathbb{P}(|Y_n - n\mathsf{q}_1| > n\epsilon) \leq \mathbb{P}\left(\left|Y_n - \sum_{k=1}^{n}\mathsf{r}_{k,1}\right| > n\epsilon/2\right) \leq 2e^{-2\frac{(n\epsilon/2)^2}{n}} = 2e^{-\frac{\epsilon^2}{2}n},$$

where the second inequality holds by Hoeffding's inequality (e.g. [12]).

It follows that

$$\begin{aligned}
\mathbb{P}(V_{n,N} \le \delta|S_N|) &\le \mathbb{P}(V_{n,N} \le \delta|S_N|, Y_n \ge (\mathsf{q}_1-\epsilon)n) + \mathbb{P}(Y_n < (\mathsf{q}_1-\epsilon)n) \\
&\le \mathbb{P}\Big(V_{n,N} \le \delta|S_N|\Big|Y_n \ge (\mathsf{q}_1-\epsilon)n\Big)\mathbb{P}(Y_n \ge (\mathsf{q}_1-\epsilon)n) + 2e^{-\frac{\epsilon^2}{2}n} \\
&\le \mathbb{P}\Big(V_{n,N} \le \delta|S_N|\Big|Y_n \ge (\mathsf{q}_1-\epsilon)n\Big) + 2e^{-\frac{\epsilon^2}{2}n}. \qquad (8)
\end{aligned}$$

Now, consider the event $\{V_{n,N} \le \delta|S_N|\}$. If all balls of colour 1 in S_N lie in a small subset, say S, of these urns then all balls of colour 1 avoid a large set $S' = S_N - S$ of urns. Hence, this set of events is equal to

$$\begin{aligned}
&\{\exists S' \subset S_N \text{ with } |S'| > (1-\delta)|S_N| : \text{ all } Y_n \text{ balls of colour 1 miss } S'\} \\
=\ &\{\exists S' \subset S_N \text{ with } |S'| = \lfloor(1-\delta)|S_N|\rfloor + 1 : \text{ all } Y_n \text{ balls of colour 1 miss } S'\} \\
=\ &\bigcup_{S' \in Q_N(1-\delta)} \{\text{all } Y_n \text{ balls of colour 1 miss } S'\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{P}\big(V_{n,N} \le \delta|S_N|\big|Y_n \ge (\mathsf{q}_1-\epsilon)n\big) &= \mathbb{P}\left(\bigcup_{S' \in Q_N(1-\delta)} \{\text{all } Y_n \text{ colour 1 balls miss } S'\}\Big|Y_n \ge (\mathsf{q}_1-\epsilon)n\right) \\
&\le \sum_{S' \in Q_N(1-\delta)} \mathbb{P}\left(\text{all } Y_n \text{ colour 1 balls miss } S'\Big|Y_n \ge (\mathsf{q}_1-\epsilon)n\right) \\
&\le \sum_{S' \in Q_N(1-\delta)} \mathbb{P}(\lceil(\mathsf{q}_1-\epsilon)n\rceil \text{ colour 1 balls all miss } S') \\
&= \sum_{S' \in Q_N(1-\delta)} \left(1 - \sum_{a \in S'} \mathsf{q}_{1,a}(N)\right)^{\lceil(\mathsf{q}_1-\epsilon)n\rceil}.
\end{aligned}$$

Putting this back into (8) gives

$$\mathbb{P}(V_{n,N} \le \delta|S_N|) \le 2e^{-\frac{\epsilon^2}{2}n} + \sum_{S' \in Q_N(1-\delta)} \left(1 - \sum_{a \in S'} \mathsf{q}_{1,a}(N)\right)^{\lceil(\mathsf{q}_1-\epsilon)n\rceil},$$

which gives the first claim of the Lemma.

For the second bound, let Y'_n be the number of balls of colour 2 chosen by time n. (It is sufficient to consider the balls of colour 2, but one could equally consider all balls of colour $c \ne 1$.) Choose n_1 such that for all $n \ge n_1$, $|n^{-1}\sum_{k=1}^n \mathsf{r}_{2,c} - \mathsf{q}_2| < \epsilon/2$. Then, as above,

$$\begin{aligned}
\mathbb{P}(V_{n,N} > \delta|S_N|, V'_{n,N} = 0) &\le \mathbb{P}(V_{n,N} > \delta|S_N|, V'_{n,N} = 0, Y'_n \ge (\mathsf{q}_2-\epsilon)n) + \mathbb{P}(Y'_n < (\mathsf{q}_2-\epsilon)n) \\
&\le \mathbb{P}(V_{n,N} > \delta|S_N|, V'_{n,N} = 0, Y'_n \ge (\mathsf{q}_2-\epsilon)n) + 2e^{-\frac{\epsilon^2}{2}n}. \qquad (9)
\end{aligned}$$

Now

$$\begin{aligned}
\{V_{n,N} > \delta|S_N|\} &\subset \{\exists S \subset S_N : |S| > \delta|S_N|, \text{ each } a \in S \text{ contains a ball of colour 1}\} \\
&= \{\exists S \subset S_N : |S| = \lfloor\delta|S_N|\rfloor + 1, \text{ each } a \in S \text{ contains a ball of colour 1}\} \\
&= \bigcup_{S \in Q_N(\delta)} \{\text{each } a \in S \text{ contains a ball of colour 1}\}.
\end{aligned}$$

Let $\{S_1, S_2, \dots S_{|Q_N(\delta)|}\}$ be an enumeration of $Q_N(\delta)$, and for $i = 1, \dots, |Q_N(\delta)|$ let

$$B_{i,n} = \{\text{each urn } a \in S_i \text{ contains a ball of colour 1}\} \cap \{Y'_n \ge (\mathsf{q}_2-\epsilon)n\}.$$

Then the first term on the right hand side of (9) is bounded by

$$\mathbb{P}\Big(\bigcup_{1 \le i \le |Q_N(\delta)|} B_{i,n}, V'_{n,N} = 0\Big) \le \mathbb{P}\Big(V'_{n,N} = 0\Big|\bigcup_{1 \le i \le |Q_N(\delta)|} B_{i,n}\Big) = \mathbb{P}\Big(V'_{n,N} = 0\Big|\bigcup_{1 \le i \le |Q_N(\delta)|} D_{i,n}\Big), \qquad (10)$$

where $D_{i,n} = B_{i,n} \setminus (\cup_{j<i} B_{i,n})$, are disjoint events. We claim that for a finite collection of disjoint events D_i (of positive probability),

$$\mathbb{P}(E|\cup_i D_i) \le \max_i \mathbb{P}(E|D_i). \tag{11}$$

To see this note that it is equivalent to

$$\frac{\sum_i \mathbb{P}(E\cap D_i)}{\sum_i \mathbb{P}(D_i)} \le \max_i \frac{\mathbb{P}(E\cap D_i)}{\mathbb{P}(D_i)}.$$

However this is true since it is easy to prove (by induction on n) that for any $x_i, y_i > 0$, $i = 1, \ldots, n$,

$$\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n y_i} \le \max_{1\le i\le n} \frac{x_i}{y_i}.$$

Applying (11) to (10) we see that (10) is bounded above by

$$\begin{aligned}\max_{i\le |Q_N(\delta)|} \mathbb{P}\Big(V'_{n,N} = 0 \Big| D_{i,n}\Big) &\le \max_{1\le i\le |Q_N(\delta)|} \mathbb{P}(\lceil(\mathsf{q}_2-\epsilon)n\rceil \text{ colour 2 balls all miss } S_i)\\ &\le \max_{1\le i\le |Q_N(\delta)|} \left(1-\sum_{a\in S_i}\mathsf{q}_{2,a}(N)\right)^{\lceil(\mathsf{q}_2-\epsilon)n\rceil},\end{aligned}$$

which completes the proof.

□

Lemma 5. *Let notation be as above, conditions as in Theorem 1 and define $M = N^{1/2+1/16}$. Then*

$$\sum_{n=M}^{\infty} P_1(n, N, N^{-1/2}) = O(1).$$

Proof. Lemma 4 showed that for n sufficiently large (not depending on N) and for $0 < \delta < 1$

$$P_1(n,N,\delta) \le 2e^{-\epsilon^2 n/2} + \sum_{S\in Q_N(1-\delta)} \left(1-\sum_{a\in S}\mathsf{q}_{1,a}(N)\right)^{\lceil \mathsf{q}_1-\epsilon\rceil n}.$$

By Assumption 2, for $S \subset S_N$ with $|S| = \lfloor(1-\delta)|S_N|\rfloor + 1$ we have

$$\mu(1-\delta) \le \sum_{a\in S}\mathsf{q}_{1,a}(N).$$

Hence,

$$P_1(n,N,\delta) \le 2e^{-\epsilon^2 n/2} + \binom{|S_N|}{\lfloor(1-\delta)|S_N|\rfloor+1}(1-\mu(1-\delta))^{\lceil \mathsf{q}_1-\epsilon\rceil n}.$$

By taking subsets of S_N, if necessary, for any $\alpha' > \alpha$ and for all N sufficiently large, there is an S_N satisfying Assumption 2 with $|S_N| = \lceil \alpha N\rceil \le \alpha' N$. Fixing such an S_N and using the bounds $\binom{n}{k} \le (ne/k)^k$ and $\delta x - 1 \le x - (\lfloor(1-\delta)x\rfloor + 1) \le \delta x$ gives (for N sufficiently large)

$$\begin{aligned}\binom{|S_N|}{\lfloor(1-\delta)|S_N|\rfloor+1} &= \binom{|S_N|}{|S_N| - (\lfloor(1-\delta)|S_N|\rfloor+1)}\\ &\le \left(\frac{|S_N|e}{\delta|S_N|-1}\right)^{\delta|S_N|}\\ &\le \left(\frac{2e}{\delta}\right)^{\delta\alpha' N}\\ &= \exp\big((1+\ln(2/\delta))\delta\alpha' N\big).\end{aligned}$$

Taking $\delta = 1/\sqrt{N} \to 0$ and $0 < \mu < 1$, for N sufficiently large, $1-\mu(1-\delta) \le 1-\mu/2 < 1$. Hence, for N and n sufficiently large (depending only on constants in the model and not depending on each other),

$$P_1(n,N,N^{-1/2}) \le 2\exp\big(-\epsilon^2 n/2\big) + \exp\big((1+\ln(2/\sqrt{N}))\alpha'\sqrt{N}\big)(1-\mu/2)^{\lceil \mathsf{q}_1-\epsilon\rceil n}.$$

Therefore, setting $\eta = (1-\mu/2)^{\lceil \mathsf{q}_1 - \epsilon \rceil} < 1$ and $\eta' = -\ln \eta > 0$,

$$\begin{aligned}\sum_{n=M}^{\infty} P_1(n, N, N^{-1/2}) &\leq O(1) + \exp((1+\ln(2/\sqrt{N}))\alpha'\sqrt{N}) \int_M^{\infty} \eta^x dx \\ &\leq O(1) + \exp((1+\ln(2/\sqrt{N}))\alpha'\sqrt{N}) \exp(-\eta' M)/\eta'.\end{aligned}$$

The result now follows from the facts that $M = N^{1/2+1/16}$ and $(1+\ln(2/\sqrt{N}))\alpha'\sqrt{N} \leq \eta'\sqrt{N}N^{1/16}$ for all N sufficiently large. □

Lemma 6. *Let notation be as above, conditions as in Theorem 1 and let $M = N^{1/2+1/16}$. Then*

$$\sum_{n=M}^{\infty} P_2(n, N, N^{-1/2}) = O(1).$$

Proof. Lemma 4 showed that for n sufficiently large (not depending on N)

$$P_2(n, N, \delta) \leq 2e^{-\epsilon^2 n/2} + \max_{S \in Q_N(\delta)} \left(1 - \sum_{a \in S} \mathsf{q}_{2,a}(N)\right)^{\lceil \mathsf{q}_2 - \epsilon \rceil n}.$$

For any S_N satisfying Assumption 2, $\mu\alpha \frac{|S|}{|S_N|} \leq \sum_{a \in S} \mathsf{q}_{2,a}(N)$ so for all $S \in Q_N(\delta)$ and sufficiently large N,

$$1 - \sum_{a \in S} \mathsf{q}_{2,a}(N) \leq 1 - \mu\alpha \frac{\delta|S_N|+1}{|S_N|} \leq 1 - \mu\alpha\frac{\delta}{2} < 1.$$

We have $1 - \mu\alpha\delta/2 \leq e^{-\frac{1}{2}\mu\alpha\delta}$ so that

$$\left(1 - \sum_{a \in S} \mathsf{q}_{2,a}(N)\right)^{\lceil \mathsf{q}_2 - \epsilon \rceil n} \leq (1 - \mu\alpha\delta/2)^{\lceil \mathsf{q}_2 - \epsilon \rceil n} \leq \exp(-\frac{1}{2}\mu\alpha \lceil \mathsf{q}_2 - \epsilon \rceil n\delta).$$

Therefore for all N and n sufficiently large (depending only on constants in the model and not depending on each other),

$$P_2(n, N, \delta) \leq 2\exp(-\epsilon^2 n/2) + \exp\big(-\tfrac{1}{2}\mu\alpha\lceil \mathsf{q}_2 - \epsilon \rceil n\delta\big).$$

It follows that

$$\begin{aligned}\sum_{n=M}^{\infty} P_2(n, N, N^{-1/2}) &\leq O(1) + \int_M^{\infty} \exp\left(-\tfrac{1}{2}\mu\alpha N^{-\frac{1}{2}}(\mathsf{q}_2 - \epsilon)x\right) dx \\ &= O(1) + \frac{2N^{\frac{1}{2}}}{\mu\alpha(\mathsf{q}_2 - \epsilon)} \exp\big(-\tfrac{1}{2}\mu\alpha N^{-\frac{1}{2}}(\mathsf{q}_2 - \epsilon)M\big) \\ &= O(1) + O\left(N^{1/2}\exp\big(-cN^{1/2+1/16-1/2}\big)\right) \\ &= O(1) + O\left(N^{1/2}e^{-cN^{1/16}}\right)\end{aligned}$$

for some constant $c > 0$. Therefore this term is $O(1)$ as required. □

4.3. **Proof of Theorem 1.** Recall that $W_N(n)$ is the number of collisions of balls of different colour when n balls have been put in urns. Then $\mathbb{P}(Z_N > n) = \mathbb{P}(W_N(n) = 0)$ and we wish to compute

$$\mathbb{E}[Z_N] = \sum_{n=0}^{\infty} \mathbb{P}(Z_N > n) = \sum_{n=0}^{\infty} \mathbb{P}(W_N(n) = 0).$$

Therefore

$$\sum_{n=0}^{M} \mathbb{P}(W_N(n) = 0) \leq \mathbb{E}[Z_N] \leq \sum_{n=0}^{M} \mathbb{P}(W_N(n) = 0) + \sum_{n>M} \mathbb{P}(Z_N > n).$$

The final term is $O(1)$ by Lemmas 5 and 6.

The first term is equal to

$$\sum_{n=0}^{M} e^{-\lambda_N(n)} + \sum_{n=0}^{M} \left(\mathbb{P}(W_N(n) = 0) - e^{-\lambda_N(n)} \right).$$

By Theorem 2 and equation (7), the second term here is at most

$$2 \sum_{n=0}^{M} \mathrm{Err}_N(n) = O(N^{1/4}).$$

Now, by Lemma 1 we have $|\lambda_N(n) - n^2 A_N/2| = O(n/N)$. Hence, for some constant $c > 0$,

$$\sum_{n=0}^{M} e^{-n^2 A_N/2 - cn/N} \le \sum_{n=0}^{M} e^{-\lambda_N(n)} \le \sum_{n=0}^{M} e^{-n^2 A_N/2 + cn/N}.$$

For $0 \le n \le M$ we have $e^{cn/N} \le e^{cM/N} = e^{c/N^{1/2-1/16}}$. Further, since $e^x \le 1 + 2x$ for $0 \le x \le 1.25$, we have $e^{c/N^{1/2-1/16}} \le 1 + 2c/N^{1/2-1/16}$ for N sufficiently large. Similarly $e^{-c/N^{1/2-1/16}} \le 1 - c/N^{1/2-1/16}$ using $1 - x \le e^{-x}$. Hence,

$$\sum_{n=0}^{M} e^{-\lambda_N(n)} = \left(1 + O(N^{-\frac{7}{16}})\right) \sum_{n=0}^{M} e^{-n^2 A_N/2}.$$

Now,

$$\int_0^M e^{-x^2 A_N/2} dx \le \sum_{n=0}^{M} e^{-n^2 A_N/2} \le 1 + \int_0^M e^{-x^2 A_N/2} dx.$$

We write the integral as

$$\int_0^\infty e^{-x^2 A_N/2} dx - \int_M^\infty e^{-x^2 A_N/2} dx.$$

As is well-known,

$$\int_0^\infty e^{-x^2 A_N/2} dx = \tfrac{1}{2}\sqrt{\pi 2/A_N} = \sqrt{\pi/(2A_N)}$$

which provides the leading term of Theorem 1.

Since $A_N \ge f/N$,

$$\int_M^\infty e^{-x^2 A_N/2} dx \le \int_M^\infty e^{-fx^2/(2N)} dx \le \int_M^\infty e^{-fMx/(2N)} dx = \frac{2N}{fM} e^{-fM^2/(2N)}.$$

This is $O(1)$ since $M = N^{1/2+1/16}$.

Putting it all together we have

$$\begin{aligned}
\mathbb{E}[Z_N] &= \sum_{n=0}^{M} \mathbb{P}(W_N(n) = 0) + O(1) \\
&= \sum_{n=0}^{M} e^{-\lambda_N(n)} + O(N^{\frac{1}{4}}) \\
&= \left(1 + O(N^{-\frac{7}{16}})\right) \sum_{n=0}^{M} e^{-\frac{1}{2} n^2 A_N} + O(N^{\frac{1}{4}}) \\
&= \left(1 + O(N^{-\frac{7}{16}})\right) \left[\sqrt{\frac{\pi}{2A_N}} + O(1) \right] + O(N^{\frac{1}{4}}) \\
&= \left(1 + O(N^{-\frac{7}{16}})\right) \sqrt{\frac{\pi}{2A_N}} + O(N^{\frac{1}{4}}).
\end{aligned}$$

Finally, since $\sqrt{1/A_N} = O(N^{1/2})$ it follows that

$$\left(1 + O(N^{-\frac{7}{16}})\right) \sqrt{\frac{\pi}{2A_N}} = \sqrt{\frac{\pi}{2A_N}} + O(N^{1/16}).$$

This completes the proof. □

Note that with more work one could presumably obtain a tighter error bound, but this is sufficient for the applications in [9, 10].

5. Applications to the DLP

Recall the discrete logarithm problem from the introduction. There are a number of algorithms to solve the DLP in an interval, but recent work of Galbraith and Ruprai [8] and Galbraith, Pollard and Ruprai [10] has shown that variants of an algorithm due to Gaudry and Schost [11] can be used to give good results. For background on the Gaudry-Schost algorithm and its analysis we refer to [8, 11].

These algorithms perform pseudorandom walks in certain sets. Recall that an instance of the DLP consists of a pair (g, h) of group elements such that $h = g^a$ for some integer a that we wish to compute. A pseudorandom walk is a sequence of group elements $x_1, x_2, \ldots$. A walk is called "tame" if each element is of the form $x_i = g^{a_i}$ where the integers a_i are known to the algorithm. A walk is called "wild" if each element is of the form $x_j = hg^{b_j}$ where the integers b_j are known to the algorithm. A collision in the walk therefore gives an equation of the form $g^{a_i} = hg^{b_j}$, from which one solves the problem as $h = g^{a_i - b_j}$. The expected number of group operations performed by the algorithm depends on the expected number of samples from the group until a tame and wild walk collide. We think of the "tame" walks as being colour 1 and the "wild" walks as being colour 2 (sometimes there is another type of wild walk, in which case a third colour is needed). The group elements are thought of as urns, and if a walk visits a group element then we think of a ball of the corresponding colour being placed in the corresponding urn.

The subclass of models relevant to the above algorithm is as follows. Let $C \in \{2, 3\}$ be the number of colours and let $\mathsf{q}_1 = \mathsf{q}_2 = \cdots = \mathsf{q}_C = 1/C$ (with $b_{k,c}$ satisfying the required property).

Let $\alpha \in (0, 1)$ and let $\{M_N\}_{N \in \mathbb{N}}$ be an $\mathbb{N}$-valued sequence satisfying $M_N < N$ and $p_N = M_N/N$. Suppose that for each colour $1 \leq c \leq C$ and each N there is a subset $S_c(N)$ of the N urns such that $|S_c(N)| = M_N$ and that for each c, a, N

$$\mathsf{q}_{c,a}(N) = \begin{cases} p_N, & a \in S_c(N) \\ 0, & \text{otherwise.} \end{cases}$$

For this class of models, since

$$\sum_{a=1}^{N} \mathsf{q}_{c,a}\mathsf{q}_{c',a} = |S_c(N) \cap S_{c'}(N)|p_N^2,$$

we have

$$A_N = \frac{p_N^2}{C^2} \sum_{c=1}^{C} \sum_{c'=1}^{C} |S_c(N) \cap S_{c'}(N)| = \frac{2p_N^2}{C^2} \sum_{c=1}^{C} \sum_{c'>c} |S_c(N) \cap S_{c'}(N)|.$$

Therefore,

$$\sqrt{\frac{\pi}{2A_N}} = \frac{C\sqrt{\pi}}{2p_N} \left(\sum_{c=1}^{C-1} \sum_{c'=c+1}^{C} |S_c(N) \cap S_{c'}(N)| \right)^{-\frac{1}{2}}. \tag{12}$$

Suppose further that there exist colours $c, c' \neq c$ such that $|S_c(N) \cap S_{c'}(N)| \geq \alpha N$. It follows immediately from our main result (taking $S_N = S_c(N) \cap S_{c'}(N)$ for the specific colours c and c' mentioned above) that the expected time of the algorithm is (12) to leading order.

More relevant for this paper is that Galbraith and Ruprai [9] have used the Gaudry-Schost algorithm to solve the discrete logarithm problem in an interval taking advantage of equivalence classes under inversion. This variant leads to non-uniform distributions on the group elements sampled, and the distributions depend on whether one is running a tame or wild walk. The analysis of the algorithm requires the main result of our paper. We now explain how Theorem 3 of [9] is derived from this result. First, we give the statement of it.

Theorem 3. *Let $N \in \mathbb{N}$ and $0 \leq A \leq N/2$. Suppose we have an unlimited number of balls of two colours, red and blue, and N urns. Suppose we alternately*[1] *choose balls of each colour and put them in random urns,*

[1]When running the algorithm on a serial computer one can arrange the computation so that tame and wild walks take alternate steps. This corresponds to alternate choices of colours. However, in practice one often considers distributed or parallel implementations and in this setting it is much more realistic to assume that balls are coloured with probability 1/2.

independently of previous urn selections. Red balls are assigned to the N urns uniformly. Blue balls are assigned to the N urns with the following probabilities: each urn $1 \le u \le A$ is chosen with probability $2/N$, each urn $A < u \le N - A$ is chosen with probability $1/N$, and urns $N - A < u \le N$ are used with probability 0. Then the expected number of assignments that need to be made in total before we have an urn containing two balls of different colour is $\sqrt{\pi N} + O(N^{1/4})$.

Proof. The colours are chosen alternately, so suppose the k-th ball has colour 1 when k is odd and colour 2 when k is even. In the language of Assumption 1 this means that $\mathsf{r}_{k,1} = (1-(-1)^k)/2$ and $\mathsf{r}_{k,2} = (1+(-1)^k)/2$. It follows that $\mathsf{q}_1 = \mathsf{q}_2 = 1/2$ and

$$b_{n,1} = \frac{1}{2} - \frac{1}{n}\sum_{k=1}^{n} \frac{1-(-1)^k}{2} = \frac{1}{2n}\sum_{k=1}^{n}(-1)^k \in \{0, -\frac{1}{2n}\}, \qquad \text{and } b_{n,2} = \frac{1}{2} - \frac{1}{n}\sum_{k=1}^{n} \frac{1+(-1)^k}{2} \in \{0, \frac{1}{2n}\}.$$

Hence $|b_{n,c}| \le K/n$ with $K = \frac{1}{2}$.

If colour 1 is red and colour 2 is blue, then in the language of Assumption 2, $\mathsf{q}_{1,a} = 1/N$ for $1 \le a \le N$, and

$$\mathsf{q}_{2,a} = \begin{cases} 2/N, & 1 \le a \le A \\ 1/N, & A < a < N - A \\ 0, & N - A \le a \le N. \end{cases}$$

Since $A \le N/2$ we can let $S_N = \{a \in \mathbb{Z} : 1 \le a \le N/2\}$ so that $|S_N| = \lfloor \frac{N}{2} \rfloor$ and $\alpha = 1/2 - \epsilon$. Since $\mathsf{q}_{c,a} \ge 1/N$ for all $a \in S_N$ and $c = 1, 2$, Assumption 2 holds. Hence, we can apply Theorem 1.

One computes

$$A_N = 2\frac{1}{2}\frac{1}{2}\left(A\frac{2}{N}\frac{1}{N} + (N-2A)\frac{1}{N}\frac{1}{N}\right) = \frac{1}{2N}.$$

Hence, by Theorem 1 the expected number of trials until there is a collision is

$$\sqrt{\pi N} + O(N^{1/4})$$

as claimed. □

As shown in [9] this leads to an algorithm for the DLP in an interval of length N with conjectural average case expected asymptotic complexity of $(1.36 + \epsilon)\sqrt{N}$ group operations for small $\epsilon > 0$. The factor ϵ comes from a number of practical issues with the algorithm (in particular the fact that we are using a pseudorandom walk and that the equivalence classes can lead to small cycles) and it is unclear whether ϵ can be made arbitrarily small without other aspects of the algorithm starting to dominate the running time.

6. A Paradox about Birthdays

In some variants of the discrete logarithm algorithms it happens that balls of one colour are assigned to a smaller set of possible urns than balls of another colour. One might believe that to minimise the expected time until a collision occurs, the latter colour should be sampled more than the former one. However, a consequence of our results is that, at least asymptotically, one should sample equally from both sets. Since this result is potentially counter-intuitive (at least to non-experts) and seems to be not well known, we give a formulation in terms of "birthday problems".

Suppose a conference center has two rooms, one holding a meeting of the "boys born in January club" and the other holding a meeting of "random girls" (i.e., ones whose birthdays are uniformly distributed over the whole year). We ask boys and girls to sequentially enter the lobby and and wish to maximise the chances that there is a girl and a boy of the same birthday, while minimising the number of people in the lobby. It is natural to ask boys and girls to enter the lobby in a manner so that the ratio of boys to girls converges to some fixed value (for example, one can ask a boy to enter the room with a certain fixed probability). What should this value be? Further, what is the best strategy as the number N of days in a year tends to infinity and the month of January contains roughly one twelfth of all days?

One's intuition might be that significantly more girls than boys should be brought into the lobby. On the contrary, it follows from our main result that (at least asymptotically) one should choose the same number

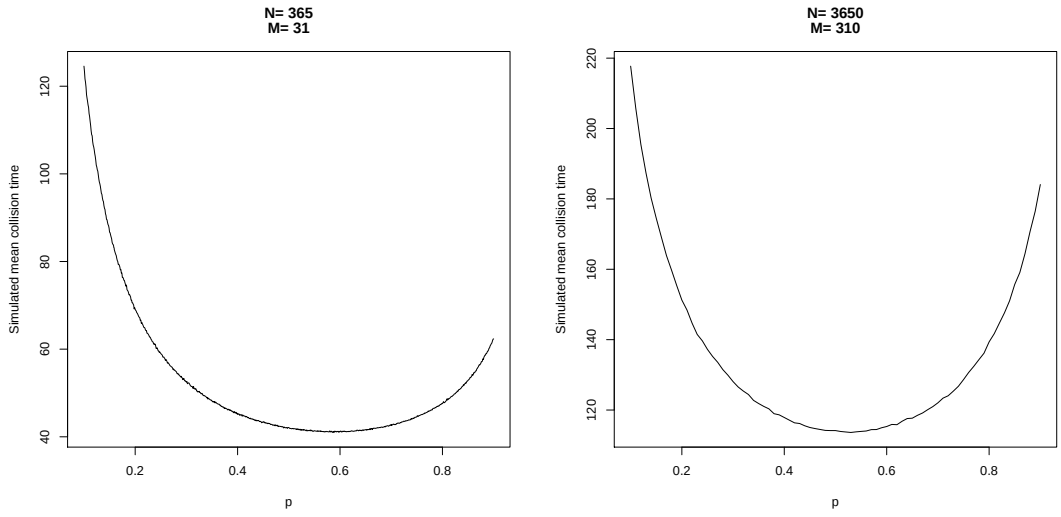

FIGURE 2. Simulations of the expected number of people chosen before a birthday collision occurs between males born in a subset of size M of the total (size N) days in a year and females born at any time of year, when $(M, N) = (31, 365)$ or $(M, N) = (310, 3650)$. The expected number is plotted as a function of $p = \mathsf{q}_2$, the probability of selecting a person from the female group, each time a person is selected.

of boys as girls. To see this, suppose boys are colour 1 and girls are colour 2. Then (assuming $12 \mid N$), $\mathsf{q}_{2,a} = 1/N$ while

$$\mathsf{q}_{1,a} = \begin{cases} 12/N, & 1 \le a \le N/12 \\ 0, & N/12 < a \le N. \end{cases}$$

The expected number of trials is of order $\sqrt{\pi/(2A_N)}$ and so one wants to maximise A_N. We have q_1 and $\mathsf{q}_2 = 1 - \mathsf{q}_1$ being the probability of choosing boys and girls respectively. The formula for A_N in this case simplifies to

$$A_N = 2\mathsf{q}_1\mathsf{q}_2(N/12)(12/N)(1/N) = 2\mathsf{q}_1\mathsf{q}_2/N.$$

It immediately follows that to maximise A_N one should choose $\mathsf{q}_1 = \mathsf{q}_2 = 1/2$, in which case the expected number of trials is asymptotically $\sqrt{\pi N}$ just as it is in the case where the boys and girls birthdays are distributed uniformly over the whole year.

Simulations (see Figure 2) suggest that, for the case $N = 365$ and January containing 31 days, the optimal choice of sampling is roughly $\mathsf{q}_2 = 0.6$. For the case $N = 3650$ and January containing 310 days, the optimal choice of sampling appears to be taking q_2 just slightly larger than 0.5.

ACKNOWLEDGEMENTS

We thank P. Diaconis for bringing the paper [4] to our attention, S. Cope for assisting with the simulations, S. Murphy for suggesting a clarification of Section 6, and the anonymous referees for several suggestions.

REFERENCES

[1] R. Arratia, L. Goldstein and L. Gordon, Two Moments Suffice for Poisson Approximations: The Chen-Stein Method, Ann. Probab., Vol. 17, No. 1 (1989) 9–25.
[2] R. Arratia, L. Goldstein and L. Gordon, Poisson Approximation and the Chen-Stein Method, Statistical Science, Vol. 5, No. 4 (1990) 403–424.
[3] M. Camarri and J. Pitman, Limit Distributions and Random Trees Derived from the Birthday Problem with Unequal Probabilities, Electronic J. Probability, Vol. 5, No. 2 (2000) 1–18.
[4] S. Chatterjee, P. Diaconis and E. Meckes, Exchangeable pairs and Poisson approximation, Electronic Encyclopedia of Probability (2004).
[5] L. H. Y. Chen, Poisson approximation for dependent trials. Ann. Probab., Vol. 3, No. 3 (1975) 534–545.

[6] A. DasGupta, The Matching, Birthday and the Strong Birthday Problem: A Contemporary Review, J. Statistical Planning and Inference, Vol. 130 (2005) 377–389.
[7] P. Flajolet, D. Gardy and L. Thimonier, Birthday paradox, coupon collectors, caching algorithms and self-organizing search, Discrete Appl. Math. 39 (1992), no. 3, 207–229.
[8] S. D. Galbraith and R. S. Ruprai, An Improvement to the Gaudry-Schost Algorithm for Multidimensional Discrete Logarithm Problems, in M. Parker (ed.), Twelfth IMA International Conference on Cryptography and Coding, Cirencester, Springer LNCS 5921 (2009) 368–382.
[9] S. D. Galbraith and R. S. Ruprai, Using Equivalence Classes to Accelerate Solving the Discrete Logarithm Problem in a Short Interval, in P. Q. Nguyen and D. Pointcheval (eds.), PKC 2010, Springer LNCS 6056 (2010) 368–383.
[10] S. D. Galbraith, J. M. Pollard and R. S. Ruprai, The Discrete Logarithm Problem in an Interval, to appear in Math. Comp.
[11] P. Gaudry and E. Schost, A low-memory parallel version of Matsuo, Chao and Tsujii's algorithm, in D. A. Buell (ed.), ANTS VI, Springer LNCS 3076 (2004) 208–222.
[12] G. R. Grimmett and D. R. Stirzaker, Probability and Random Processes 2nd Ed. Oxford University Press (1992).
[13] K. Nishimura and M. Sibuya, Occupancy with two types of balls, Ann. Inst. Statist. Math., Vol. 40, No. 1 (1988) 77–91.
[14] K. Nishimura and M. Sibuya, Probability To Meet in the Middle, J. Cryptology 2, No. 1, (1990) 13–22.
[15] J. M. Pollard, Monte Carlo methods for index computation (mod p), Math. Comp. **32** (1978), no. 143, 918–924.
[16] J. M. Pollard, Kangaroos, Monopoly and discrete logarithms, J. Crypt. **13** (2000), no. 4, 437–447.
[17] B. I. Selivanov, On waiting time in the scheme of random allocation of coloured particles, Discrete Math. Appl., Vol. 5, No. 1 (1995) 73–82.

E-mail address: S.Galbraith@math.auckland.ac.nz

Mathematics Department, The University of Auckland, Private Bag 92019 Auckland 1142 New Zealand. Phone: (+64 9) 923-87 77 FAX: (+64 9) 3737 457

E-mail address: mholmes@stat.auckland.ac.nz

Department of Statistics, The University of Auckland, Private Bag 92019 Auckland 1142 New Zealand.