The Annals of Applied Probability
2016, Vol. 26, No. 6, 3286–3318
DOI: 10.1214/16-AAP1176

COLLISION TIMES IN MULTICOLOR URN MODELS AND SEQUENTIAL GRAPH COLORING WITH APPLICATIONS TO DISCRETE LOGARITHMS

BY BHASWAR B. BHATTACHARYA

Stanford University

Consider an urn model where at each step one of q colors is sampled according to some probability distribution and a ball of that color is placed in an urn. The distribution of assigning balls to urns may depend on the color of the ball. Collisions occur when a ball is placed in an urn which already contains a ball of different color. Equivalently, this can be viewed as sequentially coloring a complete q-partite graph wherein a collision corresponds to the appearance of a monochromatic edge. Using a Poisson embedding technique, the limiting distribution of the first collision time is determined and the possible limits are explicitly described. Joint distribution of successive collision times and multi-fold collision times are also derived. The results can be used to obtain the limiting distributions of running times in various birthday problem based algorithms for solving the discrete logarithm problem, generalizing previous results which only consider expected running times. Asymptotic distributions of the time of appearance of a monochromatic edge are also obtained for other graphs.

1. Introduction. Suppose the vertices of a finite graph $G = (V, E)$, with $|V| = N$, are colored independently and uniformly at random with c colors. The probability that the resulting coloring has no monochromatic edge, that is, it is a proper coloring is $\chi_G(c)/c^N$, where $\chi_G(c)$ denotes the number of proper colorings of G using c-colors. The function χ_G is the chromatic polynomial of G, which is a central object in graph theory [18, 30]. A natural extension is to consider a general coloring distribution $\underline{p} = (p_1, p_2, \ldots, p_c)$, where the probability that a vertex is colored with color $a \in [c]$ is p_a which is independent of the colors of the other vertices, where $p_a \geq 0$, and $\sum_{a=1}^{c} p_a = 1$. Then the probability that G is properly colored is related to Stanley's generalized chromatic polynomial [19, 43]. Limit theorems for the number of monochromatic edges under the uniform coloring distribution, that is, $p_a = 1/c$ for all $a \in [c]$, was derived recently by Bhattacharya et al. [8].

When the underlying graph G is a complete graph, this reduces to the well-known birthday problem: by replacing the colors by birthdays, occurring with possibly nonuniform probabilities, the birthday problem can be seen as coloring the vertices of a complete graph independently with $c = 365$ colors. The event

Received July 2015; revised January 2016.
MSC2010 subject classifications. Primary 05C15, 60F05; secondary 94A62, 60G55.
Key words and phrases. Discrete logarithm, graph coloring, limit theorems, Poisson embedding.

that two people share the same birthday is the event of having a monochromatic edge in the colored graph. The birthday problem was generalized to the sequential setting by Camarri and Pitman [12] as follows: in a stream of people, determine the distribution of the first time that a person arrives whose birthday is the same as that of some person previously in the stream. More generally, they derived the asymptotic distribution of the first repeat time in an i.i.d. $\mathbb{P}_N$ sequence, in a limiting regime with the probability distribution $\mathbb{P}_N$ depending on a parameter $N \in \mathbb{N}$. Formally, suppose that the Nth distribution $\mathbb{P}_N$ is a *ranked discrete distribution*, $p_{N1} \geq p_{N2} \geq \cdots \geq 0$ and $\sum_{i=1}^{\infty} p_{Ni} = 1$. A sequence $\mathcal{X}_N := (X_{N1}, X_{N2}, \ldots)$ of i.i.d. random variables distributed as $\mathbb{P}_N$ is said to have a *repeat at time* t, if $X_{Nt} = X_{Ns}$, for some $s < t$. The *first repeat time*

$$R_{N1} = \inf\{t \in \mathbb{N} : r_N(t) = 1\}, \tag{1.1}$$

where $r_N(t)$ is the number of repeats in the sequence $\mathcal{X}_N$ up to time t. In other words, R_{N1} is the first time some element is observed twice in the sequence $\mathcal{X}_N$. More generally, the mth repeat time R_{Nm} of the sequence $\mathcal{X}_N$, is the minimum t such that $r_N(t) = m$, that is, the first time that m repetitions occur in the sequence $\mathcal{X}_N$.

This can also be viewed as sequentially coloring the vertices of the infinite complete graph, independently with probability $\mathbb{P}_N$, and R_{N1} is the first time that a monochromatic edge appears. Another way to rephrase this is in terms of an urn model with urns (corresponding to birthdays) indexed by $\{1, 2, \ldots\}$ and with infinitely many balls. Initially, all the urns are empty, and at every subsequent time step a ball is dropped into urn i with probability p_{Ni}, where $\sum_{i=1}^{\infty} p_{Ni} = 1$, and $p_{N1} \geq p_{N2} \cdots \geq 0$. Then R_{N1} is the first time that there are two balls in the same urn.

In the *uniform* case, where $p_{Ni} = 1/N$ for all $i \in \{1, 2, \ldots, N\}$, it is well known that for all $r \geq 0$, $R_{N1}/\sqrt{N}$ converges to the Rayleigh distribution with parameter 1. Camarri and Pitman [12] used the Poisson embedding technique and characterized the set of all possible asymptotic distributions of R_{N1} derived from any sequence of general ranked distributions. In the uniform case, Arratia et al. [2] derived the limiting distribution of the mth repeat time R_{Nm}, when $m = O(N)$.

The nonsequential version of the urn model described above is the classical occupancy scheme with infinitely many boxes, where balls are thrown independently into boxes with probability $\mathbb{P}_N$. Asymptotics for the number of boxes occupied by exactly r balls are well known [5, 26]. In a different context, Paninsky [39] used B_1, the number of boxes with 1 ball, for testing uniformity given sparsely-sampled discrete data. The Poisson embedding technique is also useful in other occupancy urn problems: Holst [27, 28] used it to derive moments of a general quota problem; Holst [29] and later Neal [38] also used these techniques to obtain limiting distributions in coupon-collector problems. For other variations of occupancy urn models and their applications, refer to [26, 32] and the references therein. For embedding Pólya-type urn schemes into continuous time Markov branching processes refer to [3, 32, 35].

1.1. *Collision times in a multicolor urn model.* A natural generalization of the birthday problem is to consider coincidences among individuals of different types, that is, in a room occupied with an equal number of boys and girls, when can one expect a boy and girl to share the same birthday. This can be viewed as an urn model with two colors, where balls are colored independently with probability $1/2$ and placed in the urns uniformly. The event of having a matching birthday is same as having an urn with balls of both the colors. This event is often referred to as a collision. For exact expressions of the number of collisions, factorial moments and other related problems, refer to Nakata [37] and the references therein. The number of collisions between two discrete distributions was also used by Batu et al. [7] for distributional property testing. Wendl [44] studied a very related problem and referred to some applications in collisions of airborne planes, celestial objects and transportation.

In this paper, we consider the sequential version of this problem, for general urn selection distributions.

DEFINITION 1.1. Consider an urn model with balls of q distinct colors (corresponding to types) indexed by $\{1, 2, \ldots, q\}$, and urns (corresponding to birthdays) indexed by $\{1, 2, \ldots\}$. Initially, all the boxes are empty, and at every subsequent time point the following steps are executed:

1. (*Color selection*) A color is $a \in [q]$ is chosen uniformly, that is, with probability $1/q$.
2. (*Urn selection*) If the color chosen is $a \in [q]$, then a ball with color a is dropped into urn i with probability p_{Ni}, where $p_{N1} \geq p_{N2} \cdots \geq 0$ and $\sum_{i=1}^{\infty} p_{Ni} = 1$, is a ranked discrete distribution.

Let C_t be the color of the ball chosen at the tth step, and Z_{Nt} be the urn to which the ball is assigned. The urn model described above is said to have a *collision at time* t, if $Z_{Ns} = Z_{Nt}$ and $C_s \neq C_t$, for some $s < t$. In other words, a collision happens when a ball is dropped in an urn which already contains a ball with a different color. Given the above process, define the *first collision time* T_{N1} *to be the first time that there exist two balls with different colors in the same urn.*

Using the Poisson embedding technique, the limiting distribution of T_{N1} can be obtained.

THEOREM 1.1. *For* $N, i \in \mathbb{N}$, *let*

$$s_N = \Big(\sum_i p_{Ni}^2\Big)^{\frac{1}{2}}, \qquad \psi_{Ni} = \frac{p_{Ni}}{s_N}. \tag{1.2}$$

Suppose that $\lim_{N\to\infty} p_{N1} = 0$, *and* $\psi_i = \lim_{N\to\infty} \psi_{Ni}$ *exists, for* $i \in \mathbb{N}$. *Then*

$$\begin{aligned} &\lim_{N\to\infty} \mathbb{P}(s_N T_{N1} > r) \\ &\quad = e^{-\frac{1}{2}(\frac{q-1}{q})r^2\cdot(1-\sum_i \psi_i^2)} \prod_i e^{-(\frac{q-1}{q})\psi_i r}\big(q - (q-1)e^{-\psi_i \frac{r}{q}}\big). \end{aligned} \tag{1.3}$$

Conversely, if there exist positive constants $c_N \to 0$ and d_N such that the distribution of $c_N(T_{N1} - d_N)$ has a nondegenerate weak limit as $N \to \infty$, then $p_{N1} \to 0$ and limits ψ_i exist as before. So the weak limit is just a rescaling of that described in (1.3), with $c_N/s_N \to \alpha$ for some $0 < \alpha < \infty$, and $c_N d_N \to 0$.

If the process described above is continued after the first collision time T_{N1}, more collisions occur. Recall that a collision corresponds to a ball being dropped in an urn which already contains a ball with a different color.

DEFINITION 1.2. For $m \geq 1$, let T_{Nm} be the time of the mth collision, that is,

$$T_{Nm} = \inf\left\{ t \in \mathbb{N} : \sum_{i=1}^{t} K_t = m \right\},$$

where K_t is the indicator variable which is 1 if and only if the urn model described above has a collision at time t.

From the continuous time embedding of the process, the joint convergence of the collision times can be obtained.

THEOREM 1.2. *Suppose $\lim_{N\to\infty} p_{N1} = 0$, $\psi_i = \lim_{N\to\infty} \psi_{Ni}$ exists for each $i \in \mathbb{N}$, and s_N as in (1.2). Then there is the convergence of m-dimensional distributions*

$$(s_N T_{N1}, s_N T_{N2}, \ldots, s_N T_{Nm}) \xrightarrow{\mathcal{D}} (\eta_1, \eta_2, \ldots, \eta_m),$$

where $0 < \eta_1 < \eta_2 < \cdots$ are the arrival times of a process $\mathcal{M}$, which is the superposition of independent point processes B^*, $B_1^{-L_1}$, $B_2^{-L_2}$, $\ldots$, where:

- *B^* is a Poisson process on $[0, \infty)$ of rate $(1 - \sum_i \psi_i^2)t \cdot (1 - \frac{1}{q})$ at time t.*
- *For each $i \in \mathbb{N}$, B_i is the superposition of q independent Poisson processes*

  $$\{B_i^1(t)\}_{t\geq 0}, \{B_i^2(t)\}_{t\geq 0}, \ldots, \{B_i^q(t)\}_{t\geq 0}$$

  on $[0, \infty)$ of rate ψ_i/q. Finally, $B_i^{-L_i}$ is the process B_i with its first $L_i := B_i(T_i')$ points removed, where T_i' is the last arrival time in B_i before $T_i = \inf\{t \geq 0 : B_i^a(t) > 0$ and $B_i^b(t) > 0$ for some, $a \neq b\}$.

The time T_i' defined above is the last arrival time when all points of B_i have the same color. Therefore, removing the first $L_i := B_i(T_i')$ points ensures that any subsequent arrival in B_i corresponds to a collision in the urn labelled i.

The urn model described in Definition 1.1 can be generalized further by considering nonuniform color selection and letting the probability of selecting an urn to depend on the color selected.

DEFINITION 1.3. Consider an urn model with balls of q distinct colors indexed by $\{1, 2, \ldots, q\}$, and urns indexed by $\{1, 2, \ldots\}$. Initially, all the boxes are empty, and at every time instance the following steps are executed:

1. (*Nonuniform color selection*) A color is chosen with probability distribution $\mathbf{c} = (c_1, c_2, \ldots c_q)$, that is, the probability of selecting the color $a \in [q]$ is c_a, where $c_a > 0$ and $\sum_{a=1}^{q} c_a = 1$.
2. (*Nonuniform urn selection*) If the color chosen is $a \in [q]$, then a ball with color a is dropped into urn i with probability $p_{Ni,a}$, where $\sum_{i=1}^{\infty} p_{Ni,a} = 1$, for all $a \in [q]$.

As in Definition 1.1, denote by T_{N1} the first time there exist two balls with different colors in the same urn.

THEOREM 1.3. *For $a \in [q]$, and $N, i \in \mathbb{N}$, let*

$$s_N^2 = \sum_i \Bigg(\sum_a c_a p_{Ni,a} \Bigg)^2, \qquad \psi_{Ni,a} = \frac{p_{Ni,a}}{s_N}.$$

Suppose that $\lim_{N\to\infty} \max_i p_{Ni,a} = 0$ and $\psi_{i,a} = \lim_{N\to\infty} \psi_{Ni,a}$ exists, for all $a \in [q]$ and $i \in \mathbb{N}$. Moreover, assume that $\phi_a = \lim_{N\to\infty} \sum_i \psi_{Ni,a}^2$ exists for all $a \in [q]$. Then

$$\begin{aligned} &\lim_{N\to\infty} \mathbb{P}(s_N T_{N1} > r) \\ &\quad = e^{-(1-\beta)\frac{r^2}{2}} \prod_i e^{-r\sum_{a=1}^{q} c_a \psi_{i,a}} \Bigg(1 + \sum_{a=1}^{q} e^{\psi_{i,a} c_a r} - q \Bigg), \end{aligned} \tag{1.4}$$

where $\beta = \sum_{a=1}^{q} c_a^2 \phi_a + \sum_i \sum_{a\neq b} c_a c_b \psi_{i,a} \psi_{i,b}$.

Theorem 1.1 is a special case of the above theorem when $c_a = 1/q$ and $p_{Ni,a} = p_{Ni}$, for all $a \in [q]$. Another special case was considered by Selivanov [42], where only Rayleigh distributions were obtained as limits.[1] Recently, Galbraith and Holmes [22] considered a variant of the urn model in Definition 1.3,

[1] Selivanov ([42], Theorem 4.1) claims that $s_N T_{N1}$ converges to a Rayleigh distribution, whenever $\sum_i p_{Ni}^2 \to 0$ and $p_{N1}(\sum_i p_{Ni}^2)^{-\frac{1}{2}} < c$, for some constant c. However, the second condition is vacuously true for all distributions, for any $c > 1$ (since $p_{Ni}^2 \le \sum_i p_{Ni}^2$, for all $i \ge 1$, implies that $p_{N1}^2 \le \sum_i p_{Ni}^2$). This implies that $s_N T_{N1}$ converges to a Rayleigh distribution, whenever $\sum_i p_{Ni}^2 \to 0$. However, Theorem 1.1 shows that this is clearly incorrect, since the conditions $p_{N1} \to 0$ and $\sum_i p_{Ni}^2 \to 0$ are equivalent (see Examples 2.2 and 2.3 for specific counterexamples). A possible fix to Selivanov's condition is to assume that $p_{N1}(\sum_i p_{Ni}^2)^{-\frac{1}{2}} \to 0$. This would imply that $\psi_i = \lim_{N\to\infty} \psi_{Ni} = 0$, for all $i \ge 1$, and by (1.3), $\lim_{N\to\infty} \mathbb{P}(s_N T_{N1} > r) = e^{-\frac{r^2}{4}}$, for $q = 2$.

where the color selection probabilities change with time, and used the Chen–Stein method to determine the expected first collision time. Theorem 1.3 extends these results and characterizes the different limiting distributions that may arise. Moreover, this general theorem can be used to find the asymptotic distributions of the running times for a class of algorithms for solving the discrete logarithm problem (DLP) that requires generalizations of the birthday problem (details in Section 1.3 and Section 5).

1.2. *Sequential graph coloring*. The repeat time R_{N1} of Pitman and Camarri [12] is the first time when a monochromatic edge appears while sequentially coloring the vertices of the (infinite) complete graph, independently with probability distribution $\mathbb{P}_N$. The collision time T_{N1} in the urn model defined in the previous section is the first time a monochromatic edge appears while sequentially coloring the (infinite) complete q-partite graph, where at every step one of the q partite sets is chosen uniformly at random, and a vertex in that set is colored independently with probability $\mathbb{P}_N$. Similar questions can be asked for any sequence of naturally growing graphs, which motivate us to formulate the following general problem.

Let $\mathcal{G} := (G_t)_{t\geq 1}$ be a deterministic sequence graphs $G_t = (V_t, E_t)$, with $V_{t+1} \subset V_t$, $|V_{t+1}| = |V_t| + 1$, and $E_t \subseteq E_{t+1}$. For $N \geq 1$, consider the following sequential coloring scheme:[2]

- Every vertex in V_1 is colored independently with a ranked discrete probability distribution $\mathbb{P}_N$.
- For $t \geq 2$, the new vertex $v \in V_t \setminus V_{t-1}$ is colored with $\mathbb{P}_N$:

$$\mathbb{P}(\text{the vertex } v \text{ has color } i \in \mathbb{N}) = p_{Ni},$$

  independent of the color all the other vertices.[3] Define the first *collision time* $T^{\mathcal{G}}_{N1}$ to be the first index s when a monochromatic edge $(u, v) \in E_s$ appears.

Note that this general framework includes the repeat time R_{N1} defined in (1.1) (take $G_t = K_t$ the complete graph on t vertices), and the collision time T_{N1} (G_t is a complete q-partite graph on t vertices with the added randomness that at every step one of the q partite sets is chosen uniformly at random).

A popular model for evolving random graphs is the *preferential attachment* (PA) model, introduced in a seminal paper by Barabási and Albert [4]. It builds on the paradigm that new vertices are attached to those already present with probability

[2]More formally, we have a triangular array of growing graphs $((G_{Ns}))_{s\geq 1}$, whose vertices are colored independently with probability distribution $\mathbb{P}_N$. For notational simplicity, the process is only described for a sequence of growing graphs $(G_t)_{t\geq 1}$, and in all the examples considered this simplification suffices.

[3]This should not be confused with the color of the ball in the urn model, described in the previous section. The urn model corresponds to coloring a complete q-partite graph, and the color of a ball corresponds to which of the q sets the vertex belongs to.

proportional to their degree. This model enjoys many properties observed in social networks and other real world networks: the power law distribution of vertex degrees, a small diameter, and a small average degree [10, 11]. For every fixed integer $m \geq 2$, the PA(m) model is formally defined as follows:

The graph sequence grows one vertex at a time, and at the tth step the graph G_m^t is an undirected graph on the vertex set $V := [t]$ defined inductively as follows. G_t^1 consists of a single vertex with m self-loops. For all $t > 1$, G_m^t is built from G_m^{t-1} by adding a new node labelled t together with m edges $e_1^t = (t, v_1), \ldots, e_m^t = (t, v_m)$ inserted one after the other in this order. Let $G_{m,i-1}^t$ denote the graph right before the edge e_i^t is added. Let $M_i = \sum_{v \in V} d_{G_{m,i-1}^t}(v)$ be the sum of the degrees of all the nodes in $G_{m,i-1}^t$. The endpoint v_i is selected randomly such that $v_i = u$ with probability $d_{G_{m,i-1}^t}(u)/(M_i + 1)$, except for t that is selected with probability $(d_{G_{m,i-1}^t}(t) + 1)/(M_i + 1)$.

Note that the graph G_m^t can have loops and multiple edges. However, it forms a vanishing fraction of the total edges and for the coloring problem it suffices to consider the underlying simple graph [to be denoted by $S(G_m^t)$].

By taking $\mathcal{G} = (S(G_m^t))_{t \geq 1}$, define $T_{N1}^{\mathrm{PA}(m)}$ to be the first time there is a monochromatic edge in the sequential coloring of $\mathcal{G}$. Using the Stein's method for Poisson approximation, the following theorem can be proved.

THEOREM 1.4. *Let s_N be as in (1.2) and $\lim_{N\to\infty} p_{N1} \to 0$, as $N \to \infty$. Then $s_N^2 T_{N1}^{\mathrm{PA}(m)} \xrightarrow{\mathcal{D}} \mathrm{Exp}(m)$, the exponential distribution with parameter m.*

The asymptotics for collision times can also be studied for any deterministic sequence of graphs which grow naturally one vertex at a time. This is demonstrated for the infinite path: take $\mathcal{Z} = (P_t)_{t \geq 1}$, where P_t is the path with t vertices, and define $T_{N,m}^{\mathcal{Z}}$ to be the first time there exists a monochromatic path with m vertices while sequentially coloring $\mathcal{Z}$. As in Theorem 1.4, the limit distribution of $T_{N,m}^{\mathcal{Z}}$ can be proved (see Theorem 6.2).

1.3. *Applications to the discrete logarithm problem.* The discrete logarithm problem (DLP) in a finite group G is as follows: given $g, h \in G$ find an integer a such that $h = g^a$. Due to its presumed computational difficulty, the problem figures prominently in various cryptosystems, including the Diffie–Hellman key exchange, El Gamal system and elliptic curve cryptosystems. The best algorithms to solve the discrete logarithm problem in a general group originate in the seminal work of Pollard [40, 41]. A standard variant of the classical Pollard Rho algorithm for finding discrete logarithms can be described using a Markov chain on the cycle. The running time of the algorithm is the collision time of the Markov chain that is, the first time the chain visits a state that was already visited. Several years later, Kim et al. [33] finally proved the widely believed $\Theta(\sqrt{|G|})$ collision time for this walk.

The discrete logarithm problem in an interval asks: given $g, h \in G$ and an integer N find an integer a, if it exists, such that $h = g^a$ and $0 \le a < N$. The DLP in an interval can be solved using the baby-step-giant-step algorithm in $\sqrt{2}N$ group operations and storage of $O(\sqrt{N})$ group elements. The Pollard kangaroo method [41] was designed to solve the DLP in an interval using a constant number of group elements of storage. Using distinguished points a heuristic average case expected complexity of essentially $2\sqrt{N}$ group operations and low storage can be obtained. Montenegro and Tetali [36] gave a more rigorous analysis of the kangaroo method. Recently, Galbraith et al. [23] used a 4-kangaroo method, instead of the usual two, to obtain a heuristic average case expected running time of $(1.715 + o(1))\sqrt{N}$.

There has been several recent work extending and improving Pollard's algorithms for variants of the discrete logarithm problem which require generalizations of the birthday problem [21, 23, 24]. Gaudry and Schost [25] presented one of the first birthday problem based methods for solving the DLP in an interval. The algorithm is based on the collision time of 2 independent pseudo-random walks. A *tame walk* is a sequence of points $\{g^{a_i}\}_{i \ge 1}$ where $a_i \in T$ and a *wild walk* is a sequence of points $g^{b_i} = hg^{a_i}$ with $b_i \in W$, where $T, W \subseteq \{1, 2, \ldots, |G|\}$ are the tame and wild sets, respectively. When the same element is visited by two different types of walk, there is a tame-wild collision giving an equation of the form $g^{a_i} = hg^{b_j}$, and the DLP is solved as $h = g^{a_i - b_j}$. Therefore, the running time of the algorithm is the time required until a tame and wild walk collide. The average expected running time of this algorithm is $2.08\sqrt{N}$ group operations on a serial computer. Recently, Galbraith et al. [23] proposed a four-set Gaudry–Schost algorithm with heuristic average case expected running time of $(1.661 + o(1))\sqrt{N}$. Later, modifying the Gaudry–Schost algorithm and using a variant of the birthday paradox, Galbraith and Ruprai [24] proposed an improvement in groups in which inversions are faster than general group operations, as in elliptic curves. This algorithm will be referred to as the accelerated Gaudry–Schost algorithm, and has a heuristic average case expected running time of approximately $1.36\sqrt{N}$ group operations.

In the analyses of all such algorithms the quantity used to compare the running times is the expectation of the tame-wild collision time, averaged over all problem instances. However, in the light of the above theorems, the asymptotic distribution of the running time of all such algorithms can be obtained, under the assumption that the pseudo-random walks performed by the algorithms are sufficiently random and their running times can be analyzed by an idealized birthday problem involving the tame-wild collision. This is derived for the Gaudry–Schost algorithm (Theorem 5.1) and the accelerated Gaudry–Schost algorithm of Galbraith and Ruprai (Theorem 5.2). To the best of our knowledge, these are the first known results about the limiting distributions of the running times of these algorithms. Though these results are based on some heuristic assumptions, they give considerable insight about the dependence between the running times and the complexity of the problem instance.

1.4. *Organization of the paper.* The paper is organized as follows: The proofs of Theorem 1.1 and Theorem 1.2, and examples are given in Section 2. An analogous limit theorem for the m-fold collision time is proved in Section 3. In Section 4, the generalized urn model is considered and the proof of Theorem 1.3 is presented. The asymptotic distributions of the running times of algorithms for the discrete logarithm problem are proved in Section 5. The limiting distributions of the collision times for the preferential attachment model and the infinite path are derived in Section 6.

2. Proofs of Theorems 1.1 and 1.2. In this section, limiting distributions of collision times in the urn model described in Definition 1.1 are derived.

2.1. *Proof of Theorem 1.1.* Let $\mathcal{P}$ be a homogeneous Poisson process on $\mathcal{R} := [0,\infty) \times [0,1]$ of rate 1 per unit area, with points $\{(S_1, W_1), (S_2, W_2), \ldots\}$, where $0 < S_1 < S_2 <$ are the points of a homogeneous Poisson process on $[0,\infty)$ of rate 1 per unit length, and $W_1, W_2, \ldots$ are i.i.d. Unif$(0,1)$. Let $\mathcal{R}_t = [0,t] \times [0,1]$ and $\mathcal{P}(t)$ be the restriction of $\mathcal{P}$ to $\mathcal{R}_t$.

- Color the points in $\mathcal{P}$ independently with one of q colors, $\{1, 2, \ldots, q\}$ with probability $1/q$, that is,
$$\mathbb{P}\big((S_i, W_i) \in \mathcal{P} \text{ has color } a \in [q]\big) = 1/q,$$
independently for every point in $\mathcal{P}$. For $a \in [q]$, denote by $\mathcal{P}^a$ the subsets of $\mathcal{P}$ colored $a \in [q]$. By the marking theorem [34], $\mathcal{P}^1, \mathcal{P}^2, \ldots, \mathcal{P}^q$ are independent Poisson process each with rate $1/q$ on $\mathcal{R}$.
- For $N \geq 1$, partition $[0,1]$ into intervals $J_{N1}, J_{N2}, \ldots,$ such that the length of J_{Ni} is p_{Ni} (see Figure 1). For $t \geq 0$ and $a \in [q]$, let
$$\mathcal{P}_{Ni} = \mathcal{P} \cap [0,\infty) \times J_{Ni} \quad \text{and} \quad \mathcal{P}^a_{Ni} = \mathcal{P}^a \cap [0,\infty) \times J_{Ni}. \tag{2.1}$$
Clearly, $\mathcal{P}_{N1}, \mathcal{P}_{N2}, \ldots$ are independent Poisson processes with rates p_{N1}, $p_{N2}, \ldots,$ respectively; and for $a \in [q]$, $\mathcal{P}^a_{N1}, \mathcal{P}^a_{N2}, \ldots$ are independent Poisson processes with rates $p_{N1}/q, p_{N2}/q, \ldots,$ respectively.

The collision time T_{Nm} (Definition 1.2) can be described in terms of the above process: let C_j be the color of the point (S_j, W_j) and $Z_{Nj} = \sum_i i\mathbf{1}\{W_j \in J_{Ni}\}$. The sequence $\{(C_j, Z_{Nj})\}_{j\geq 1}$ in the discrete time model corresponds to the color of the jth ball and the urn to which the jth ball is assigned. For $j \in \mathbb{N}$, define
$$K_j := \mathbf{1}\big\{\exists n \in \mathbb{N} \text{ with } Z_{Nj} = Z_{Nj'} = n \text{ and } C_{j'} \neq C_j, \text{ for some } j' < j\big\},$$
the indicator that there is a collision at the jth step. The mth collision time (recall Definition 1.2) is defined as
$$T_{Nm} \stackrel{D}{=} \inf\left\{ j \in \mathbb{N} : \sum_{i=1}^{j} K_i = m \right\}. \tag{2.2}$$

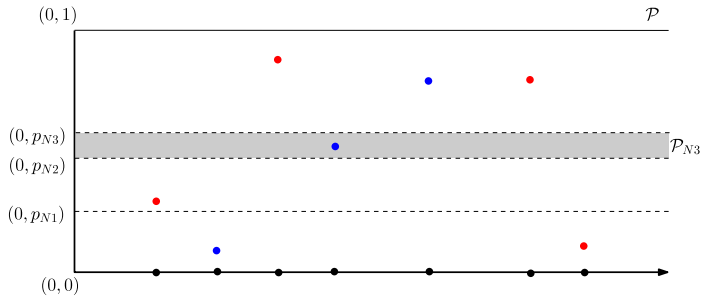

FIG. 1. *A schematic of the Poisson embedding for $q = 2$ colors. Points are colored red or blue with probability $\frac{1}{2}$.*

In particular, the first collision time T_{N1} (defined as in Definition 1.1) is

$$\inf\{j \in \mathbb{N} : \exists n \in \mathbb{N} \text{ with } Z_{Nj} = Z_{Nj'} = n \text{ and } C_{j'} \neq C_j, \text{ for some } j' < j\}.$$

LEMMA 2.1. *Let $\tau_{Nm} = \inf\{t \in \mathbb{R} : |\mathcal{P}(t)| \geq T_{Nm}\}$, where T_{Nm} is as defined in (2.2). Then $\frac{T_{Nm}}{\tau_{Nm}} \xrightarrow{P} 1$, whenever $p_{N1} \to 0$, as $N \to \infty$.*

PROOF. By the strong law of large numbers $|\mathcal{P}(t)|/t$ converges almost surely to 1 as $t \to \infty$. Therefore, it suffices to show τ_{N1} converges in probability to infinity as $N \to \infty$, since by definition $|\mathcal{P}(\tau_{N1})| = T_{N1}$. This implies τ_{Nm} converges in probability to infinity, as $\tau_{Nm} \geq \tau_{N1}$, for $m \geq 1$.

By definition τ_{N1} is

$$\inf\{t \geq 0 : \exists j \in \mathbb{N} \text{ with } |\mathcal{P}^{\alpha}_{Nj}(t)| > 0, |\mathcal{P}^{\beta}_{Nj}(t)| > 0, \text{ for some } \alpha \neq \beta \in [q]\},$$

where $\mathcal{P}^{a}_{Nj}(t)$ is the restriction of $\mathcal{P}^{a}_{Nj}$ [defined in (2.1)] to $\mathcal{R}_t$, for $a \in [q]$. This implies that

$$\begin{aligned}\mathbb{P}(\tau_{N1} > t) &= \prod_i \big(q\big(1 - e^{-p_{Ni}\frac{t}{q}}\big)e^{-(q-1)p_{Ni}\frac{t}{q}} + e^{-p_{Ni}t}\big) \\ &= e^{-(\frac{q-1}{q})t} \prod_i \big(q - (q-1)e^{-p_{Ni}\frac{t}{q}}\big).\end{aligned} \tag{2.3}$$

Using $\log(q - (q-1)e^{-\frac{x}{q}}) \geq \frac{q-1}{q}(x - \frac{x^2}{2})$, for $x \geq 0$, (2.3) simplifies to

$$\big|\log \mathbb{P}(\tau_{N1} > t)\big| \leq \frac{1}{2}\left(\frac{q-1}{q}\right)t^2 \sum_i p_{Ni}^2 \leq \frac{1}{2}\left(\frac{q-1}{q}\right)t^2 p_{N1} \sum_i p_{Ni} \to 0,$$

and the result follows. □

Let $p_N = (p_{N1}, p_{N2}, \ldots)$ be the vector of probabilities. By the above lemma, to get the limiting distribution of T_{N1} it suffices to derive the limiting distribution

of τ_{N1}. From (2.3), $\log(\mathbb{P}(\tau_{N1} > t)) = g(t, p_N)$, where

$$g(r, \underline{\psi}) := \sum_i \log\{e^{-(\frac{q-1}{q})r\psi_i}(q - (q-1)e^{-\psi_i \frac{r}{q}})\}, \tag{2.4}$$

for $r \geq 0$ and a vector $\underline{\psi} = (\psi_1, \psi_2, \ldots)$.

LEMMA 2.2. *Let $\underline{\psi} = (\psi_1, \psi_2, \ldots)$ and $\psi_1 \geq \psi_2 \geq \ldots \geq 0$ and $\sum_i \psi_i^2 < \infty$. Then there exists a constant $0 < c(q) \leq 1$, depending only on q, such that for $r \in R := [0, c(q)/\psi_1)$, there exists $\{a_s\}_{s=3}^{\infty}$ nonnegative constants with*

$$g(r, \underline{\psi}) = -\frac{1}{2}\left(\frac{q-1}{q}\right)r^2 \sum_{i=1}^{\infty} \psi_i^2 + \sum_{s\geq 3}(-1)^{s+1} a_s r^s \sum_{i=1}^{\infty} \psi_i^s, \tag{2.5}$$

and the above series is absolutely convergent.

PROOF. For $c(q)$ chosen small enough, $|1 - e^{-\psi_1 r}| < 1/(q-1)$, for $r \in R$. As $\psi_i \leq \psi_1$, $|1 - e^{-\psi_i r}| < 1/(q-1)$, for $r \in R$ and all $i \in \mathbb{N}$. Now, using the expansion of $\log(1+z)$, for $|z| < 1$,

$$\begin{aligned} g(r, \underline{\psi}) &= \sum_i \left\{ -\left(\frac{q-1}{q}\right)r\psi_i + \sum_{s=1}^{\infty} \frac{(-1)^{s+1}(q-1)^s}{s}(1 - e^{-\psi_i \frac{r}{q}})^s \right\} \\ &= \sum_i \left\{ -\left(\frac{q-1}{q}\right)r\psi_i + \sum_{s=1}^{\infty} \frac{(-1)^{s+1}(q-1)^s}{s}\left(\sum_{x=1}^{\infty}(-1)^{x+1}\frac{\psi_i^x r^x}{q^x x!}\right)^s \right\} \\ &= T_1 + T_2, \end{aligned} \tag{2.6}$$

where

$$T_1 = \sum_i (q-1)\left(\sum_{x=2}^{\infty}(-1)^{x+1}\frac{\psi_i^x r^x}{q^x x!}\right),$$

and

$$T_2 = \sum_i \sum_{s=2}^{\infty} \frac{(-1)^{s+1}(q-1)^s}{s}\left(\sum_{x=1}^{\infty}(-1)^{x+1}\frac{\psi_i^x r^x}{q^x x!}\right)^s.$$

Define

$$\begin{aligned} \mathscr{S} = \sum_i \Bigg\{ &(q-1)\left(\sum_{x=2}^{\infty}\frac{\psi_i^x r^x}{q^x x!}\right) \\ &+ \sum_{s=2}^{\infty} \sum_{\gamma_1, \ldots, \gamma_s \geq 1} \frac{(q-1)^s}{s} \frac{(\psi_i r)^{\sum_{b=1}^s \gamma_b}}{q^{\sum_{b=1}^s \gamma_b} \prod_{b=1}^s \gamma_b!} \Bigg\}. \end{aligned}$$

To show that (2.6) is absolutely convergent, it suffices to show $\mathscr{S} < \infty$, whenever $r \in R$ and $C := \sum_i \psi_i^2 < \infty$. Let $\lambda = \frac{r\psi_1}{q}$, and observe,

$$\begin{aligned}\mathscr{S} &\le C(q-1)\sum_{x=2}^{\infty}\frac{\psi_1^{x-2}r^x}{q^x} + \frac{C}{\psi_1^2}\sum_{s=2}^{\infty}(q-1)^s \sum_{\gamma_1,\dots,\gamma_s\ge 1}\left(\frac{r\psi_1}{q}\right)^{\sum_{b=1}^{s}\gamma_b}\\ &\le \frac{C(q-1)r^2}{q^2}\frac{1}{1-\lambda} + \frac{C}{\psi_1^2}\sum_{s=2}^{\infty}(q-1)^s\left(\sum_{x=1}^{\infty}\lambda^x\right)^s \qquad (2.7)\\ &\le \frac{C(q-1)r^2}{q^2}\frac{1}{1-\lambda} + \frac{C}{\psi_1^2}\sum_{s=0}^{\infty}\left(\frac{\lambda(q-1)}{1-\lambda}\right)^s < \infty,\end{aligned}$$

when $r \in R$. Therefore, by expanding (2.6) further and interchanging the order of the summation using the absolute convergence, (2.5) follows. □

2.1.1. *Completing the Proof of* (1.3) *in Theorem* 1.1. Let s_N and ψ_{Ni} be as defined in the statement of the theorem. By Lemma 2.1, it suffices to obtain the limiting distribution of τ_{N1}. From (2.3), it follows

$$\mathbb{P}(s_N\tau_{N1} > r) = \prod_i e^{-(\frac{q-1}{q})\psi_{Ni}r}\big(q-(q-1)e^{-\psi_{Ni}\frac{r}{q}}\big). \qquad (2.8)$$

As $\sum_i \psi_{Ni}^2 = 1$ and $\lim_{N\to\infty}\psi_{Ni} = \psi_i$ exists for all i, by Fatou's lemma, $\sum_i \psi_i^2 < \infty$. This implies that $\lim_{i\to\infty}\psi_i = 0$. Therefore, for every fixed $r > 0$ and there exists $j(r), N(r)$ be such that for $N > N(r)$, $\psi_{Nj(r)} < c(q)/r$. Lemma 2.2 then implies

$$\begin{aligned}\mathcal{T}_N &:= \sum_{i>j(r)}\log\{e^{-(\frac{q-1}{q})\psi_{Ni}r}\big(q-(q-1)e^{-\psi_{Ni}\frac{r}{q}}\big)\}\\ &= -\frac{1}{2}\left(\frac{q-1}{q}\right)r^2\sum_{i>j(r)}\psi_{Ni}^2 + \sum_{s=3}^{\infty}(-1)^{s+1}a_s r^s\sum_{i>j(r)}\psi_{Ni}^s, \qquad (2.9)\end{aligned}$$

where $\{a_s\}_{s\ge1}$ are nonnegative constants. Note that $\lim_{N\to\infty}\sum_{i>j(r)}\psi_{Ni}^2 = 1 - \sum_{i\le j(r)}\psi_i^2$. Moreover, for any $s \ge 3$ and $i > j(r)$, $\psi_{Ni}^s \le \psi_{Nj(r)}^{s-2}\psi_{Ni}^2$ and $\sum_i \psi_{Ni}^2 = 1$. Therefore, taking limit in (2.9) as $N \to \infty$,

$$\mathcal{T}_N \to -\frac{1}{2}\left(\frac{q-1}{q}\right)r^2\left(1-\sum_{i=1}^{j(r)}\psi_i^2\right) + \sum_{i>j(r)}\log\big(q-(q-1)e^{-\psi_i\frac{r}{q}}\big). \qquad (2.10)$$

Moreover, as $N \to \infty$,

$$\begin{aligned}&\prod_{i=1}^{j(r)} e^{-(\frac{q-1}{q})\psi_{Ni}r}\big(q-(q-1)e^{-\psi_{Ni}\frac{r}{q}}\big)\\ &\qquad\to \prod_{i=1}^{j(r)} e^{-(\frac{q-1}{q})\psi_i r}\big(q-(q-1)e^{-\psi_i\frac{r}{q}}\big),\end{aligned}$$

which combined with (2.10) gives (1.3).

2.1.2. *Proof of the converse in Theorem* 1.1. The converse to (1.3) is proved using the convergence of types, and the following lemma.

LEMMA 2.3. *Let* $\alpha > 0$ *and* $\underline{\psi} := (\psi_i, i \geq 1)$ *a nonincreasing sequence of reals with* $\sum_{i=1}^{N} \psi_i^2 \leq 1$. *Then* $(\alpha, \underline{\psi})$ *can be uniquely reconstructed from the function* $r \to h(\alpha r, \underline{\psi})$ *for* $r \in [0, \infty)$, *where*

$$h(r, \underline{\psi}) := e^{-\frac{1}{2}(\frac{q-1}{q})r^2 \cdot (1 - \sum_i \psi_i^2)} \times \prod_i e^{-(\frac{q-1}{q})\psi_i r}\big(q - (q-1)e^{-\psi_i \frac{r}{q}}\big). \tag{2.11}$$

PROOF. Using (2.6), the sequence α, $\sum_{i=1}^{\infty} \psi_i^3, \sum_{i=1}^{\infty} \psi_i^4, \ldots$, can be uniquely extracted from the function $r \to h(\alpha r, \underline{\psi})$. Now, let $(J_i, i \geq 0)$ be a partition of the unit interval such that the length of J_0 is $1 - \sum_{i=1}^{\infty} \psi_i^2$ and the length of J_i is ψ_i^2, for all $i \geq 1$. Define $Z := \sum_{i=1}^{\infty} \psi_i \mathbf{1}\{U \in J_i\}$, where U is a uniform $[0, 1]$ random variable. Then $\mathbb{E}(Z^k) = \sum_{i=1}^{\infty} \psi^{k+2}$, and these moments of Z uniquely determine the distribution of Z on $[0, 1]$. Finally, it is easily seen that this distribution uniquely determines the sequence $(\psi_1, \psi_2, \ldots)$. □

The converse to (1.3) now follows using the above lemma, and by taking sub-sequential limits and an application of convergence of types (Theorem 14.2, Billingsley [9]).

2.2. *Proof of Theorem* 1.2. Let $\mathcal{P}$ be a homogeneous Poisson process on $[0, \infty) \times [0, 1]$ of rate 1 per unit area, and $\mathcal{P}_{Ni}$, $\mathcal{P}_{Ni}^a$ be as defined before, for $i \in \mathbb{N}$ and $a \in [q]$. Note that the process $\mathcal{P}_{Ni}$ is a Poisson process of rate p_{Ni}, which is the superposition of q independent Poisson processes $\mathcal{P}_{Ni}^1, \mathcal{P}_{Ni}^2 \ldots, \mathcal{P}_{Ni}^q$ each of rate p_{Ni}/q.

Define

$$F_{Ni} = \inf\{t \geq 0 : |\mathcal{P}_{Ni}^a(t)| > 0 \text{ and } |\mathcal{P}_{Ni}^b(t)| > 0, \text{ for some } a \neq b \in [q]\}.$$

Note that the process $\mathcal{P}_{Ni}$ has points $(S_1, W_1), (S_2, W_2), \ldots$, where the inter-arrival times $S_1, S_2 - S_1, \ldots$ have independent exponential distribution with mean $1/p_{Ni}$, and $W_1, W_2, \ldots$ are i.i.d. Unif$(0, 1)$. Every point of $\mathcal{P}_{Ni}$ is colored by a color $a \in [q]$ with probability $1/q$, and the set of points colored a is the process $\mathcal{P}_{Ni}^a$. Let $\mathcal{P}_{Ni}^{-L_{Ni}}$ be the process $\mathcal{P}_{Ni}$ obtained removing the first L_{Ni} points, where $L_{Ni} = |\mathcal{P}_{Ni}(F'_{Ni})|$ and F'_{Ni} is the last arrival time in $\mathcal{P}_{Ni}$ before F_{Ni}, that is, the last arrival time when all points in $\mathcal{P}_{Ni}$ are marked with the same color. Note that F'_{Ni} is distributed as $\sum_{j=1}^{W} W_j$, where W_j are i.i.d. exponential with mean $1/p_{Ni}$ and W is a geometric with parameter $1/q$, that is,

$$\mathbb{P}(W = w) = \frac{1}{q^{w-1}}\left(1 - \frac{1}{q}\right),$$

for $w \geq 1$. By conditioning on W and calculating the characteristic function, it follows that $\sum_{j=1}^{W} W_j$ has a exponential distribution with mean $r(q)/p_{Ni}$, where $r(q) = q/(q-1)$.

Now, let $\mathcal{P}_{Ni}^{-L_{Ni}}(t) := \mathcal{P}_{Ni}^{-L_{Ni}}([0,t] \times [0,1])$, and define the counting process $X_N := (X_N(t), t \geq 0)$ as

$$X_N(t) := \sum_{i=1}^{\infty} |\mathcal{P}_{Ni}^{-L_{Ni}}(t/s_N)|.$$

The above series is bounded by $|\mathcal{P}(t/s_N)|$ and so it converges. Note that $(s_N T_{N1}, s_N T_{N2}, \ldots, s_N T_{Nm})$ are the arrival times of this process. As $\frac{T_{Nm}}{\tau_{Nm}} \xrightarrow{P} 1$, for all $m \geq 1$, by the standard theory of weak convergence of point processes (Daley and Vere-Jones [14], Theorem 9.1.VI) it is enough to show that the processes X_N converge weakly to $\mathcal{M}$.

The process $\mathcal{P}_{Ni}(\cdot)$ is a homogeneous Poisson process of rate p_{Ni}, with compensator $(p_{Ni} r, r \geq 0)$. Thus, the process $(\mathcal{P}_{Ni}(\cdot/s_N), t \geq 0)$ has compensator $(\psi_{Ni} t, t \geq 0)$ and the compensator of $\mathcal{P}_{Ni}^{-L_{Ni}}(\cdot/s_N)$ is $C_{Ni}(t) = \psi_{Ni}(t - s_N F'_{Ni})_+$, where $s_N F'_{Ni}$ has a exponential distribution with mean $r(q)/\psi_{Ni}$. Consider the following three cases:

Case 1. $\lim_{N\to\infty} \psi_{N1} = 0$. For $N, i \geq 1$ let $\mathcal{F}^{Ni} := (\mathcal{F}_t^{Ni}, t \geq 0)$ be the natural filtration of $\mathcal{P}_{Ni}(\cdot/s_N)$ and let $\mathcal{F}^N$ be the smallest filtration containing $\{\mathcal{F}^{Ni} : i \geq 1\}$. Let $(C_{Ni}(t), t \geq 0)$ be the compensator of $\mathcal{P}_{Ni}^{-L_{Ni}}(\cdot/s_N)$ with respect to the filtration $\mathcal{F}^{Ni}$ and $(C_N(t), t \geq 0)$ the compensator of X_N with respect to $\mathcal{F}^N$. Thus, $C_N(t) = \sum_i C_{Ni}(t)$,

$$\mathbb{E}\big(C_N(t)\big) = \sum_i \left(e^{-\psi_{Ni} \frac{t}{r(q)}} - 1 + \psi_{Ni} \frac{t}{r(q)} \right)$$

and

$$\operatorname{Var}\big(C_N(t)\big) = \sum_i \left(1 - e^{-2\psi_{Ni} \frac{t}{r(q)}} - 2\psi_{Ni} \frac{t}{r(q)} e^{-\psi_{Ni} \frac{t}{r(q)}} \right).$$

Now, by elementary inequalities as in [12], Lemma 11, it can be shown that $\mathbb{E}(C_N(t)) \to t^2/2r(q)^2$ and $\operatorname{Var}(C_N(t)) \to 0$ for $t > 0$. This implies that X_N converges weakly to the inhomogeneous Poisson process of rate $t/r(q)$ at time t, as required.

Case 2. $\sum_i \psi_i^2 < 1$. Let $(j_N \geq 1)$ be such that $\lim_{N\to\infty} \sum_{i \leq j_N} \psi_{Ni}^2 = \sum_i \psi_i^2$. Define the process $X_N^*(t) := \sum_{i > j_N} |\mathcal{P}_{Ni}^{-L_{Ni}}(t/s_N)|$, and $X_{Ni}(t) = |\mathcal{P}_{Ni}^{-L_{Ni}}(t/s_N)|$. Clearly, X_{Ni} converges weakly to $B_i^{-L_i}$. Moreover, as in the previous case it can be shown that $X_N^*(s_N t/s_N^*)$ converges weakly to the inhomogeneous Poisson process of rate $t/r(q)$ at time t. As $(s_N^*/s_N)^2 \to 1 - \sum_i \psi_i^2$, independence then implies that

$$\big(X_N^*, X_{N1}, \ldots, X_{Nj_N}, 0, 0, \ldots\big) \xrightarrow{D} \big(B^*, B_1^{-L_1}, B_2^{-L_2}, \ldots\big).$$

Case 3. $\sum_i \psi_i^2 = 1$. Let $(j_N \geq 1)$ be a sequence with $\lim_{N\to\infty} \sum_{i\leq j_N} \psi_{Ni}^2 = 1$. Define the process $X_N^*(t)$, and $X_{Ni}(t)$ as before. Clearly, X_{Ni} converges weakly to $B_i^{-L_i}$. Now, it is easy to show that the compensator $C_N^*(t)$ of $X_N^*(s_N t/s_N^*)$ satisfies: $\mathbb{E}(C_N^*(t)) \to 0$ and $\mathrm{Var}(C_N^*(t)) \to 0$, and the result follows.

2.2.1. *Corollary for the uniform case.* Under the uniform distribution, that is, $p_{Ni} = 1/N$, for $i \in \{1, 2, \ldots, N\}$, the limiting distribution of the first collision time T_{N1} is a Rayleigh distribution.

COROLLARY 2.1. *Suppose there are $q \geq 2$ colors and $p_{Ni} = 1/N$, for $i \in [N]$, then the distribution of $T_{N1}/\sqrt{N}$ is the Rayleigh distribution with parameter $\sqrt{1 - 1/q}$, that is,*

$$\lim_{N\to\infty} \mathbb{P}(T_{N1}/\sqrt{N} > r) = e^{-\frac{1}{2}(\frac{q-1}{q})r^2}.$$

Moreover, the distribution of $T_{Nm}/\sqrt{N}$ converges to $\sqrt{\frac{q}{q-1} \cdot \chi_{2m}^2}$, where χ_{2m}^2 is a Chi-squared distribution with $2m$ degrees of freedom.

PROOF. The limiting distribution of $T_{N1}/\sqrt{N}$ follows directly from Theorem 1.1. Theorem 1.2 implies that the limiting distribution of $T_{Nm}/\sqrt{N}$ is the time of the mth arrival of an inhomogeneous Poisson process $\mathcal{P}_\lambda$ with rate $\lambda(t) = t/r(q)$, where $r(q) = q/(q-1)$. Denote the mth arrival time by κ_m, and $|\mathcal{P}_\lambda(t)|$ the number of arrivals in $\mathcal{P}_\lambda$ up to time t. Using $|\mathcal{P}_\lambda(t)| \sim \mathrm{Pois}(t^2/2r(q))$ and $\mathbb{P}(\kappa_m < t) = \mathbb{P}(|\mathcal{P}_\lambda(t)| \geq m)$, the result follows. □

2.3. *Examples.* In this section, connections of Theorem 1.1 to the famous birthday problem are discussed. Other examples involving nonuniform urn selection probabilities are also given, illustrating the generality of the above results.

EXAMPLE 2.1 (Birthday problem). The classical birthday problem asks for the minimum number of people in a room such that two of them have the same birthday with probability at least 50%. It is well known that the minimum number people for which this holds is approximately 23. In fact, the expected number of samples, chosen uniformly with replacement, required from a set of size N until some value is repeated, is asymptotically $\sqrt{\pi N/2}$. A generalization of this considers birthday coincidences among individuals of different types, that is, in a room with equal numbers of boys and girls, when can one expect a boy and girl to share the same birthday. Finding matches among different types can also be stated in terms of sampling colored balls and placing them in urns: Suppose there are N urns and two colors, and the balls are colored independently with probability $1/2$ and placed in the urns uniformly with probability $1/N$. The number of draws needed to have 2 balls with different colors in the same urn is the first time

when a boy and a girl share the same birthday, when boys and girls sequentially enter a room independently with probability $1/2$. In this case, Corollary 2.1 for $q = 2$ shows that the limiting distribution of the first collision time is a Rayleigh distribution with parameter $\sqrt{1/2}$. This implies that the expected time of the first collision is $\sqrt{\pi N}$. When $N = 365$ and the birthdays are assumed to be uniformly distributed over the year, the expected time before there is a boy and a girl with the same birthday is 34. For a detailed discussion on the birthday problem and its various generalizations and applications, refer to [1, 6, 15–17, 31] and the references therein.

The following two examples exhibit the range of distributions that can be obtained from Theorem 1.1 when the urn selection distribution is nonuniform.

EXAMPLE 2.2. Consider the probability distribution

$$p_{N1} = \frac{1}{\sqrt{N}} \quad \text{and} \quad p_{Ni} = \frac{c_N}{N}, \qquad \text{for } i \in [2, N+1], \tag{2.12}$$

where is $c_N = 1 - \frac{1}{\sqrt{N}}$, is such that $\sum_i p_{Ni} = 1$. Note that $c_N \to 1$, as $N \to \infty$, and $\psi_1 = 1/\sqrt{2}$ and $\psi_i = 0$ for all $i \in [2, N+1]$. Therefore, by Theorem 1.1,

$$\lim_{N\to\infty} \mathbb{P}(s_N T_{N1} > r) = e^{-(\frac{q-1}{q})\cdot\frac{r^2}{4}} e^{-(\frac{q-1}{q})\frac{r}{\sqrt{2}}} \big(q - (q-1)e^{-\frac{r}{q\sqrt{2}}}\big).$$

Note that in this case $\sum_{i=1}^{\infty} \psi_i^2 < 1$, so in (1.3) both the exponential term outside the product, and the terms inside the product are nonvanishing. For $q = 2$, the limiting distribution has the following simpler form:

$$\lim_{N\to\infty} \mathbb{P}(s_N T_{N1} > r) = e^{-\frac{r^2}{8}} \big(2e^{-\frac{r}{2\sqrt{2}}} - e^{-\frac{r}{\sqrt{2}}}\big).$$

EXAMPLE 2.3. Consider the following nonuniform urn selection distribution:

$$p_{N1} = \frac{1}{\log N} \quad \text{and} \quad p_{Ni} = \frac{c_N}{N}, \qquad \text{for } i \in [2, N+1], \tag{2.13}$$

where is $c_N = 1 - \frac{1}{\log N}$, is such that $\sum_i p_{Ni} = 1$. Note that $c_N \to 1$, as $N \to \infty$, and in this case $\psi_1 = 1$ and $\psi_i = 0$ for all $i \in [2, N+1]$. Therefore,

$$\lim_{N\to\infty} \mathbb{P}(s_N T_{N1} > r) = e^{-(\frac{q-1}{q})r} \big(q - (q-1)e^{-\frac{r}{q}}\big).$$

Note that in this case $\sum_{i=1}^{\infty} \psi_i^2 = 1$, and so the exponential term in (1.3) outside the product vanishes. For $q = 2$, the limiting distribution simplifies to

$$\lim_{N\to\infty} \mathbb{P}(s_N T_{N1} > r) = 2e^{-\frac{r}{2}} - e^{-r}.$$

3. Limiting distribution of the *m*-fold collision time. Recall the urn model in Definition 1.1, and analogous to the first collision time T_{N1}, define the *m-fold collision time* $T_{N1,m}$ as the first time there exists an urn with m balls of color a and m balls of color b, for some $a \neq b \in [q]$.

The next theorem gives the asymptotic distribution of $T_{N1,m}$. Calculations are similar to those in the proof of Theorem 1.1, and some details are omitted.

THEOREM 3.1. *Let $m \geq 1$ be a fixed positive integer, and*

$$s_N^{(2m)} = \Big(\sum_i p_{Ni}^{2m}\Big)^{\frac{1}{2m}} \quad \text{and} \quad \psi_{Ni}^{(2m)} = \frac{p_{Ni}}{s_N^{(2m)}}.$$

Suppose $\lim_{N\to\infty} p_{N1} = 0$, and $\psi_i^{(2m)} := \lim_{N\to\infty} \psi_{Ni}^{(2m)}$ exist for each $i \in \mathbb{N}$. Then for $r \geq 0$,

$$\begin{aligned}\lim_{N\to\infty} \mathbb{P}\big(s_N^{(2m)} T_{N1,m} > r\big) &= e^{-\beta_m r^{2m}} \prod_i \Big\{ h_m\Big(\frac{\psi_i^{(2m)} r}{q}\Big)^{q-1} \Big[q - (q-1) h_m\Big(\frac{\psi_i^{(2m)} r}{q}\Big)\Big]\Big\}, \end{aligned} \tag{3.1}$$

where $\beta_m = (1 - \sum_{i=1}^{\infty} (\psi_i^{(2m)})^{2m}) \frac{(q-1)}{2q^{2m-1}(m!)^q}$ and $h_m(x) = \sum_{y=0}^{m-1} e^{-x} \frac{x^y}{y!}$.

PROOF. We consider the same embedding of the process as before: let $\mathcal{P}$ be a homogeneous Poisson process on $[0,\infty) \times [0,1]$ of rate 1 per unit area, and $\mathcal{P}_{Ni}$ and $\mathcal{P}_{Ni}^a$ be as defined in (2.1). The m-fold collision time can be defined as in (2.2) in terms of continuous-time process, and let $\tau_{N1,m} = \inf\{t : |\mathcal{P}(t)| \geq T_{N1,m}\}$. By the strong law of large numbers, $|\mathcal{P}(t)|/t$ converges almost surely to 1 as $t \to \infty$. As $\tau_{N1,m} \geq \tau_{N1}$, for $m \geq 1$, by Lemma 2.1, $\lim_{N\to\infty} \tau_{N1,m} = \infty$ whenever $p_{N1} \to 0$. This implies $\frac{T_{N1,m}}{\tau_{N1,m}} \xrightarrow{P} 1$, whenever $p_{N1} \to 0$, as $|\mathcal{P}(\tau_{N1,m})| = T_{N1,m}$.

Therefore, it suffices to derive the asymptotic distribution of $\tau_{N1,m}$. By definition, $\tau_{N1,m}$ is

$$\inf\{t \geq 0 : \exists j \in \mathbb{N} \text{ with } |\mathcal{P}_{Nj}^{\alpha}(t)| \geq m \text{ and } |\mathcal{P}_{Nj}^{\beta}(t)| \geq m, \text{ for } \alpha \neq \beta \in [q]\}.$$

This implies

$$\begin{aligned}\mathbb{P}(\tau_{N1,m} > r) &= \prod_i \Big[q \cdot h_m\Big(\frac{p_{Ni} r}{q}\Big)^{q-1} \Big(1 - h_m\Big(\frac{p_{Ni} r}{q}\Big)\Big) + h_m\Big(\frac{p_{Ni} r}{q}\Big)^q \Big] \\ &= \prod_i \Big\{ h_m\Big(\frac{p_{Ni} r}{q}\Big)^{q-1} \Big[q - (q-1) h_m\Big(\frac{p_{Ni} r}{q}\Big)\Big]\Big\}, \end{aligned} \tag{3.2}$$

where $h_m(x) = \sum_{y=0}^{m-1} e^{-x}\frac{x^y}{y!}$. Note that

$$\begin{aligned}
&(q-1)\log h_m(x) + \log\bigl(q-(q-1)h_m(x)\bigr)\\
&\quad= (q-1)\log\bigl(1-\bigl(1-h_m(x)\bigr)\bigr) + \log\bigl(1+(q-1)\bigl(1-h_m(x)\bigr)\bigr)\\
&\quad= -\sum_{k=2}^{\infty}\frac{(q-1)+(-1)^k(q-1)^k}{k}\cdot\bigl(1-h_m(x)\bigr)^k \qquad (3.3)\\
&\quad= -\sum_{k=2}^{\infty}\frac{(q-1)+(-1)^k(q-1)^k}{k}\cdot\Biggl(\sum_{y=m}^{\infty}e^{-x}\frac{(x)^y}{y!}\Biggr)^k\\
&\quad= -\frac{q(q-1)}{2}\cdot\frac{x^{2m}}{(m!)^2} + \sum_{k=2m+1}^{\infty}(-1)^{k+1}a_k x^k.
\end{aligned}$$

The interchange of the different summations is justified by the absolute convergence of the series, which can be proved by arguments similar to those in Lemma 2.2. Combining (3.2) and (3.3), we get

$$\begin{aligned}
&\log\mathbb{P}\bigl(s_N^{(2m)}\tau_{N1,m} > r\bigr)\\
&\quad= -\frac{q(q-1)}{2}\cdot\frac{r^{2m}}{q^{2m}(m!)^q} + \sum_{k=2m+1}^{\infty}(-1)^{k+1}a_k r^k q^{-k}\sum_i\bigl(\psi_{Ni}^{(2m)}\bigr)^k,
\end{aligned}$$

since $\sum_i(\psi_{Ni}^{(2m)})^{2m} = 1$. Finally, $(\psi_{Ni}^{(2m)})^k \to (\psi_i^{(2m)})^k$ for $k > 2m$, and using absolute convergence, $\lim_{N\to\infty}\log\mathbb{P}(s_N^{(2m)}\tau_{N1,m} > t)$ simplifies to

$$-\beta_m t^{2m} + \sum_i\Biggl\{h_m\Bigl(\frac{\psi_i^{(2m)}t}{q}\Bigr)^{q-1}\Biggl[q-(q-1)h_m\Bigl(\frac{\psi_i^{(2m)}t}{q}\Bigr)\Biggr]\Biggr\},$$

where β_m is as defined (3.1). This completes the proof of the result. □

4. Generalizing the urn model: Proof of Theorem 1.3. The proof of Theorem 1.3 presented below is similar to that of Theorem 1.1 but requires more careful calculations. To this end, recall the urn model from Definition 1.3 with nonuniform color and nonuniform urn selection probabilities. As before, the collision time T_{N1} is the first time that there exist two balls with different colors in the same urn.

4.1. *Proof of Theorem 1.3.* Let $\mathbb{P}_N$ be a ranked discrete distribution and $\mathbf{c} = (c_1, c_2, \ldots, c_q)$ be the coloring distribution as in Definition 1.3. Let $\mathcal{P}$ be a homogeneous Poisson process on $\mathcal{S} := [0,\infty)\times[0,1]$ of rate 1 per unit area, with points $\{(S_1, W_1), (S_2, W_2), \ldots\}$, where $0 < S_1 < S_2 < \cdots$ are the points of a homogeneous Poisson process on $[0,\infty)$ of rate 1 per unit length, and $W_1, W_2, \ldots$ are i.i.d. Unif(0, 1). Let $\mathcal{R}_t = [0,t]\times[0,1]$ and $\mathcal{P}(t)$ be the restriction of $\mathcal{P}$ to $\mathcal{R}_t$.

- Color the points in $\mathcal{P}$ independently with one of q colors, $\{1, 2, \ldots, q\}$ as follows:

$$\mathbb{P}\big((S_i, W_i) \in \mathcal{P} \text{ has color } a \in [q]\big) = c_a,$$

  independently over the points in $\mathcal{P}$. For $a \in [q]$, denote by $\mathcal{P}^a$ the subsets of $\mathcal{P}$ colored $a \in [q]$. By the marking theorem [34], $\mathcal{P}^1, \mathcal{P}^2, \ldots, \mathcal{P}^q$ are independent Poisson process with rates c_a on $\mathcal{R}$, respectively.
- For each $a \in [q]$ and $N \geq 1$, partition $[0, 1]$ into intervals $J_{N1,a}, J_{N2,a}, \ldots,$ such that the length of $J_{Ni,a}$ is $p_{Ni,a}$. For $t \geq 0$, let

$$\mathcal{P}^a_{Ni} = \mathcal{P}^a \cap [0, \infty) \times J_{Ni,a}. \tag{4.1}$$

  Clearly, $\mathcal{P}^a_{N1}, \mathcal{P}^a_{N2}, \ldots$ are independent Poisson processes with rates $c_a p_{N1}$, $c_a p_{N2}, \ldots,$ respectively.

Recall the definition of the first collision time T_{N1} for the urn model in Definition 1.3. It can be described in terms of the above process as follows: let C_j be the color of the point (S_j, W_j) and $Z_{Nj} = \sum_i i\mathbf{1}\{W_j \in J_{Ni,C_j}\}$. The sequence $\{(C_j, Z_{Nj})\}_{j\geq 1}$ in the discrete time model corresponds to the color of the jth ball and the urn to which the jth ball is assigned. In particular, the first collision time T_{N1} is

$$\inf\big\{j \in \mathbb{N} : \exists n \in \mathbb{N} \text{ with } Z_{Nj} = Z_{Nj'} = n \text{ and } C_{j'} \neq C_j, \text{ for some } j' < j\big\}.$$

LEMMA 4.1. *Let* $\tau_{N1} = \inf\{t : |\mathcal{P}(t)| \geq T_{N1}\}$. *Then* $\frac{T_{N1}}{\tau_{N1}} \xrightarrow{P} 1$, *whenever* $\lim_{N\to\infty} \max_i p_{Ni,a} = 0$, *for all* $a \in [q]$.

PROOF. By the strong law of large numbers, $|\mathcal{P}(t)|/t$ converges almost surely to 1 as $t \to \infty$. Therefore, it suffices to show τ_{N1} converges in probability to infinity as $N \to \infty$, since $|\mathcal{P}(\tau_{N1})| = T_{N1}$.

By definition, τ_{N1} is

$$\inf\big\{t \geq 0 : \exists i \in \mathbb{N} \text{ with } \big|\mathcal{P}^\alpha_{Ni}(t)\big| > 0, \big|\mathcal{P}^\beta_{Ni}(t)\big| > 0, \text{ for some } \alpha \neq \beta \in [q]\big\},$$

where $\mathcal{P}^a_{Nj}(t)$ is the restriction of $\mathcal{P}^a_{Nj}$ [defined in (4.1)] to $\mathcal{R}_t$, for $a \in [q]$. This implies that

$$\begin{aligned}
\mathbb{P}(\tau_{N1} > t) &= \prod_i \left(\sum_{a=1}^q (1 - e^{-p_{Ni,a}c_a t}) \prod_{b\neq a} e^{-p_{Ni,b}c_b t} + \prod_{a=1}^q e^{-p_{Ni,a}c_a t} \right) \\
&= \prod_i \left(\sum_{a=1}^q e^{-c_a t \sum_{b\neq a} p_{Ni,b}} - (q-1) e^{-t\sum_{a=1}^q p_{Ni,a}c_a} \right) \qquad (4.2) \\
&= e^{-t\sum_i \sum_{a=1}^q c_a p_{Ni,a}} \prod_i \left(1 + \sum_{a=1}^q e^{p_{Ni,a}c_a t} - q \right).
\end{aligned}$$

As $\max_i p_{Ni,a} \to 0$, choose $N > N(t,a)$ so that $p_{Ni,a} < \frac{1}{c_a t}\log(1+\frac{1}{q})$ for all $i \in \mathbb{N}$. Therefore, for $N \geq \max_a N(t,a)$ and some constant $C > 0$

$$\begin{aligned}
&\log \mathbb{P}(\tau_{N1} > t) \\
&\quad \geq \sum_i \left\{ -t \sum_{a=1}^{q} p_{Ni,a} c_a + \left(\sum_{a=1}^{q} e^{p_{Ni,a} c_a t} - q \right) - \frac{1}{2} \left(\sum_{a=1}^{q} e^{p_{Ni,a} c_a t} - q \right)^2 \right\} \\
&\quad \geq \frac{t^2}{2} \sum_i \sum_a c_a^2 p_{Ni,a}^2 - \frac{Ct^2}{2} \sum_i \left(\sum_a c_a p_{Ni,a} \right)^2 .
\end{aligned}$$

The first inequality uses $\log(1+x) \geq x - \frac{x^2}{2}$, for $|x| < 1$, and the second uses: (a) $e^x \geq 1 + x + \frac{x^2}{2}$ on the first exponential term and (b) $e^x \leq 1 + Cx$, for some constant $C := C(q)$ when $|x| \leq \log(1+1/q)$ on the second exponential term.

Now, as $\sum_i p_{Ni,b} = 1$, for all $b \in [q]$,

$$\sum_i \left(\sum_a p_{Ni,a} \right)^2 \leq \left(\max_{i \in \mathbb{N}} \sum_a p_{Ni,a} \right) \sum_a \sum_i p_{Ni,a} = q \max_{i \in \mathbb{N}} \sum_a p_{Ni,a}.$$

Therefore,

$$\begin{aligned}
\left| \log \mathbb{P}(\tau_{N1} > t) \right| &\leq \frac{Ct^2}{2} \sum_i \left(\sum_a c_a p_{Ni,a} \right)^2 \leq \frac{Ct^2}{2} \max_{a \in [q]} c_a^2 \sum_i \left(\sum_a p_{Ni,a} \right)^2 \\
&\leq \frac{Cqt^2}{2} \max_{a \in [q]} c_a^2 \max_{i \in \mathbb{N}} \sum_a p_{Ni,a} \\
&\to 0,
\end{aligned}$$

and the result follows. □

With $s_N = (\sum_i (\sum_{a=1}^{q} c_a p_{Ni,a})^2)^{\frac{1}{2}}$, $\psi_{Ni,a}$ as defined in the statement of the theorem, and (4.2),

$$\begin{aligned}
\log \mathbb{P}(s_N T_{N1} > r) &= \sum_i \log \left\{ e^{-r \sum_i \sum_{a=1}^{q} c_a \psi_{Ni,a}} \left(1 + \sum_{a=1}^{q} e^{r c_a \psi_{Ni,a}} - q \right) \right\} \\
&:= g(r, \underline{\psi}_{N,1}, \ldots, \underline{\psi}_{N,q}),
\end{aligned} \tag{4.3}$$

where $\underline{\psi}_{N,a} = (\psi_{N1,a}, \psi_{N2,a}, \ldots)$, for $a \in [q]$ and the function g is defined in (A.1).

As $\sum_i \psi_{Ni,a}^2 < \infty$ and $\lim_{N \to \infty} \psi_{Ni,a} = \psi_{i,a}$ exists for all i and $a \in [q]$, by Fatou's lemma $\sum_i \psi_{i,a}^2 < \infty$. Therefore, for $a \in [q]$, $\lim_i \psi_{i,a} = 0$, and for $r > 0$ there exists $N(r)$, $j(r)$ such that

$$\psi_{Nj(r),a} < \frac{1}{r} \log(1 + 1/q) \qquad \text{for all } N > N(r).$$

Let A and B be the functions defined in Lemma A.3. Define

$$\Gamma = \bigcup_{k=1}^{\infty} \left\{ (\gamma_1, \gamma_2, \ldots, \gamma_k) \in \mathbb{N}^k : \sum_{b=1}^{k} \gamma_b \geq 3 \right\}.$$

For $(\gamma_1, \gamma_2, \ldots, \gamma_k) \in \Gamma$ and $i > j(r)$,

$$\begin{aligned} \prod_{b=1}^{k} \sum_{a=1}^{q} c_a^{\gamma_b} \psi_{Ni,a}^{\gamma_b} &\leq \left(\sum_{a=1}^{q} c_a \psi_{Ni,a} \right)^{\sum_{b=1}^{a} \gamma_b} \\ &\leq \left(\max_{a \in [q]} \max_{i > j(r)} c_a \psi_{Ni,a} \right)^{\sum_{b=1}^{a} \gamma_b - 2} \left(\sum_{a=1}^{q} c_a \psi_{Ni,a} \right)^2. \end{aligned}$$

Using this and Lemma A.3, it follows that

$$\begin{aligned} &\lim_{N \to \infty} A(r, \underline{\psi}_{N,a}, \underline{\psi}_{N,a}, \ldots, \underline{\psi}_{N,q}) = A(r, \underline{\psi}_1, \underline{\psi}_2, \ldots, \underline{\psi}_q), \\ &\lim_{N \to \infty} B(r, \underline{\psi}_{N,a}, \underline{\psi}_{N,a}, \ldots, \underline{\psi}_{N,q}) = B(r, \underline{\psi}_1, \underline{\psi}_2, \ldots, \underline{\psi}_q). \end{aligned} \tag{4.4}$$

Finally, by assumption $\lim_{N \to \infty} \sum_i \psi_{Ni,a}^2 = \phi_a$ exists for all $a \in [q]$, and hence, as $N \to \infty$,

$$\begin{aligned} \left(\sum_i \left(\sum_{a=1}^{q} c_a \psi_{Ni,a} \right)^2 - \sum_{a=1}^{q} c_a^2 \sum_i \psi_{Ni,a}^2 \right) &\to 1 - \sum_{a=1}^{q} c_a^2 \left(\lim_{N \to \infty} \sum_i \psi_{Ni,a}^2 \right) \\ &= 1 - \sum_{a=1}^{q} c_a^2 \phi_a. \end{aligned} \tag{4.5}$$

Combining equations (4.4) and (4.5) and using Lemma A.3, the result follows.

4.2. *Useful corollaries and examples continued.* A special case of Theorem 1.3 is to consider the case where $p_{Ni,a} = p_{Ni}$, for all $a \in [q]$, and a general coloring distribution $\mathbf{c} = (c_1, c_2, \ldots, c_q)$. This simplifies (1.4) to the following.

COROLLARY 4.1. *For $a \in [q]$, and $N, i \in \mathbb{N}$, let $s_N^2 = \sum_i p_{Ni}^2$, and $\psi_{ni} = \frac{p_{Ni}}{s_N}$. Suppose that $\lim_{N \to \infty} p_{N1} = 0$ and $\psi_i = \lim_{n \to \infty} \psi_{Ni}$ exists, for each $i \in \mathbb{N}$. Then, as $N \to \infty$,*

$$\mathbb{P}(s_N T_{N1} > r) \to e^{-\frac{1}{2}(1 - \sum_i \psi_i^2)(1 - \sum_{a=1}^{q} c_a^2) r^2} \prod_i e^{-r\psi_i} \left(1 + \sum_{a=1}^{q} e^{\psi_i c_a r} - q \right).$$

The main application of Theorem 1.3 is in deriving the limiting distributions of the running times of algorithms for the discrete logarithm problem (DLP) in an interval. To this end, assume $q = 2$ colors and consider 2 discrete ranked distributions $p_{N1} \geq p_{N2} \geq \cdots$ and $q_{N1} \geq q_{N2} \geq \cdots$; where at each step one of the two

colors is chosen with probability $1/2$. If color 1 is chosen then a ball of color 1 is put in the ith urn with probability p_{Ni}, otherwise a ball of color 2 is put in the ith urn with probability q_{Ni}.

COROLLARY 4.2. *For $a \in [q]$, and $N, i \in \mathbb{N}$, let*

$$s_N = \frac{1}{2}\left(\sum_i (p_{Ni} + q_{Ni})^2\right)^{\frac{1}{2}}, \qquad \psi_{Ni} = \frac{p_{Ni}}{s_N} \quad \text{and} \quad \theta_{Ni} = \frac{q_{Ni}}{s_N}. \tag{4.6}$$

Suppose $p_{N1} \to 0$, $q_{N1} \to 0$, as $N \to \infty$, and $\lim_{N\to\infty} \psi_{Ni} = \lim_{N\to\infty} \theta_{Ni} = 0$, for all $i \in \mathbb{N}$, and $\phi_1 = \lim_{N\to\infty} \sum_i \psi_{Ni}^2$, $\phi_2 = \lim_{N\to\infty} \sum_i \theta_{Ni}^2$ exists. Then

$$\lim_{N\to\infty} \mathbb{P}(s_N T_{N1} > r) = e^{-(1-\frac{1}{4}\sum_{a=1}^2 \phi_a)\frac{r^2}{2}}. \tag{4.7}$$

The setup of Theorem 1.3 is very general and it can be used in various applications. However, when the urn selection distribution depends on the color of the ball, sometimes the scaling in Theorem 1.3 may not give a nontrivial limiting distribution as indicated in the following example.

EXAMPLE 4.1. Let $p_{N1} \geq p_{N2} \geq \cdots$ be the probability distribution (2.13), and consider the following process: Every time choose one of two colors independently with probability $1/2$; if color 1 is chosen, then with probability p_{Ni} a ball colored 1 goes to the ith urn, otherwise color 2 is chosen and a ball with that color goes to the ith urn with probability $\frac{1}{N+1}$. Let T_{N1} be the first collision time. In this case, $s_N \log N \to 1$, where s_N is as defined in (4.6), and Theorem 1.3 gives $s_N T_N = T_N / \log N$ converges to infinity in probability. However, in this case, it can be easily shown that

$$\lim_{N\to\infty} \mathbb{P}(T_N/\sqrt{N} \geq r) = e^{-\frac{r^2}{4}},$$

the Rayleigh distribution with parameter $\sqrt{2}$.

5. Algorithms for the discrete logarithm problem: Limiting distribution of running times. The central idea of the Gaudry–Schost (GS) algorithm, as well as, the kangaroo algorithm of Pollard is based on the collision time of 2 independent pseudo-random walks. Let g and h be the DLP instance, with $h = g^a$ for some integer $-N/2 \leq a \leq N/2$, where N is the size of the interval. The cyclic group G generated by g will often be described in terms of the exponent space. Define the *tame set* $T = [-N/2, N/2]$ and the *wild set* $W = a + T = \{a + b : b \in [-N/2, N/2]\}$. A *tame walk* is a sequence of points $\{g^{a_i}\}_{i\geq 1}$ where $a_i \in T$ and a *wild walk* is a sequence of points $g^{b_i} = hg^{a_i}$ with $b_i \in W$. Each walk proceeds until a distinguished point is hit. This distinguished point is then stored on a server, together with the corresponding exponent and a flag indicating which sort of walk

it was. When the same distinguished point is visited by two different types of walk, there is a tame-wild collision giving an equation of the form $g^{a_i} = hg^{b_j}$, and the DLP is solved as $h = g^{a_i - b_j}$.

The actual GS algorithm is much more complicated and although the starting point of the pseudo-random walk will be random, inherently the rest of the steps are not random, and only a heuristic running time can be derived. Experimental evidence show that the pseudo-random walks get close enough to a random selection. Therefore, it is standard in the literature to assume that when N is sufficiently large the pseudo-random walks performed by the algorithm is sufficiently random, and the running time can be analyzed by an idealized birthday problem involving the tame-wild collision. Throughout the paper, we work with this assumption, and refer to the running times of these algorithms as *idealized* running times. Then, identifying the tame walks as being color 1 and the wild walks as being color 2, and the group elements as urns, the idealized running time of the GS algorithm is precisely when two balls of different colors are placed in the same urn.

Generally, only the expectation of the collision time is used to quantify the performance of these algorithms, using the birthday problem. In the following theorem, the limiting distribution of the idealized running time of the GS algorithm for any problem instance is determined. It is assumed that the elements from T and W are sampled with probability $1/2$ each, which means that at each step the two colors are chosen with probability $1/2$ each. This is quite a realistic assumption as in practice one often considers distributed or parallel implementations of the algorithm [22, 24].

THEOREM 5.1. *Given an instance (g, h) of the DLP with $h = g^{xN}$, where $x \in [-1/2, 1/2]$, the limiting distribution of the idealized running time $T_N^{(x)}$ of the GS algorithm is*

$$\lim_{N\to\infty} \mathbb{P}\big(T_N^{(x)} > r\sqrt{N}\big) = e^{-(\frac{1-|x|}{2})\frac{r^2}{2}}.$$

PROOF. By symmetry, it suffices to consider $0 \le x < 1/2$. This implies that $|T \cap W| = (1-x)N$. Define $p_{Ni} = 1/N$, for $i \in T$, and $q_{Ni} = 1/N$, for $i \in W$. Then by (4.6)

$$s_N = \sqrt{\frac{1 - x/2}{N}} \quad \text{and} \quad \lim_{N\to\infty} \psi_{Ni} = \lim_{N\to\infty} \theta_{Ni} = 0.$$

Moreover, $\phi_1 = \phi_2 = \lim_N \sum_i \psi_{Ni}^2 = \lim_{N\to\infty} \sum_i \theta_{Ni}^2 = \frac{1}{1-x/2}$. Therefore, applying Theorem 1.3 the result follows. □

REMARK 5.1. Theorem 5.1 shows $T_N^{(x)}/\sqrt{N}$ converges to a Rayleigh distribution with parameter $(\frac{2}{1-x})^{\frac{1}{2}}$. Therefore, it is expected that in the limit

$\mathbb{E}(T_N^{(x)})/\sqrt{N} \to (1-x)^{-\frac{1}{2}}\sqrt{\pi}$. (This can be made rigorous by showing the uniform integrability of the sequence $T_N^{(x)}/\sqrt{N}$, for example, by arguing that the second moment of $T_N^{(x)}/\sqrt{N}$ is bounded.) Assuming x is uniformly distributed over $[-1/2, 1/2]$ gives, $2\int_0^{\frac{1}{2}}(1-x)^{-\frac{1}{2}}\sqrt{\pi N} = (4-2\sqrt{2})\sqrt{\pi N} \approx 2.08\sqrt{N}$. This is the leading term of the expected heuristic running time of the GS algorithm averaged over all problem instances, which was proved earlier in [24], Theorem 2, using the birthday paradox.

5.1. *Accelerated Gaudry–Schost (AGS) algorithm.* In groups where computing h^{-1} for any group element h is much faster than a general group operation, the GS algorithm can be greatly accelerated by performing random walks in sets of equivalence classes corresponding to the tame and wild sets. As before, let N, g and h be given such that $4|N$, $h = g^a$ and $-N/2 \le a \le N/2$. Define the tame and wild sets (as sets of equivalence classes) by

$$\tilde{T} = \{\{b, -b\} : b \in [-N/2, N/2]\},$$
$$\tilde{W} = \{\{a+b, -(a+b)\} : b \in [-N/4, N/4]\}.$$

Note that $|\tilde{T}| = 1 + N/2 \approx N/2$. The algorithm samples alternately from $\tilde{T}$ and $\tilde{W}$ with probability $1/2$.

THEOREM 5.2. *Given an instance (g, h) of the DLP with $h = g^{xN}$, where $x \in [-1/2, 1/2]$, the limiting distribution of the idealized running time $T_N^{(x)}$ of the AGS algorithm of Galbraith and Ruprai* [24] *is*

$$\lim_{N\to\infty} \mathbb{P}\big(T_N^{(x)} > r\sqrt{N}\big) = \begin{cases} e^{-\frac{r^2}{2}}, & \text{if } |x| < 1/4, \\ e^{-(3-4|x|)\frac{r^2}{4}}, & \text{if } 1/4 \le |x| \le 1/2. \end{cases} \tag{5.1}$$

PROOF. Let (g, h) be the DLP instance with $h = g^{xN}$, and $T_N^{(x)}$ the idealized running time of the AGS algorithm of Galbraith and Ruprai [24]. The analysis has two cases:

Case 1. $0 \le x < 1/4$. In this case $\tilde{W} \subseteq \tilde{T}$ and the algorithm samples from $\tilde{T}$ and $\tilde{W}$ alternately and uniformly. This is equivalent to sampling uniformly from $[0, N/2]$ with probability $1/2$, and sampling an element b from $[0, N/2]$ with probability $4/N$, for $0 \le b < N/4 - |x|N$, and probability $2/N$, for $N/4 - |x|N \le b \le |x|N + N/4$, with probability $1/2$. In this case,

$$s_N = \sqrt{\frac{10-8x}{4N}} \quad \text{and} \quad \lim_{N\to\infty} \psi_{Ni} = \lim_{N\to\infty} \theta_{Ni} = 0.$$

Moreover, $\phi_1 = \frac{8}{10-8x}$ and $\phi_2 = \frac{4(4-8x)}{10-8x}$. Therefore, applying Theorem 1.3 it follows

$$\lim_{N\to\infty} \mathbb{P}\left(T_N > r\sqrt{\frac{4N}{10-8x}}\right) = e^{-(\frac{1}{5-4x})r^2}.$$

Case 2. $1/4 \le x \le 1/2$. In this case, $|\tilde{T} \cap \tilde{W}| = N(3/4 - |x|)$ (here $|\tilde{T} \cap \tilde{W}|$ refers to the number of equivalence classes in the intersection). The algorithm samples uniformly between the two sets $\tilde{T}$ and $\tilde{W}$, where $|\tilde{T}| = |\tilde{W}| \approx N/2$, and as in the proof of Theorem 5.1 the limiting distribution of the idealized running time can be obtained. □

REMARK 5.2. Note that the limit (5.1) is a distribution function for every $x \in [-1/2, 1/2]$, where x is the unknown exponent in the DLP problem. Generally, x is assumed to be uniformly distributed over $[-1/2, 1/2]$ and the expected running time (as in Remark 5.1, to make this rigorous, one has to argue that the sequence $T_N^{(x)}/\sqrt{N}$ is uniformly integrable) will be

$$2\sqrt{N}\left(\int_0^{1/4}\int_0^{\infty} e^{-\frac{r^2}{2}}\,dr\,dx + \int_{1/4}^{1/2}\int_0^{\infty} e^{-(3-4|x|)\frac{r^2}{4}}\,dr\,dx\right)$$
$$= (5\sqrt{2}/4 - 1)\sqrt{\pi N} \approx 1.36\sqrt{N}.$$

This is the leading term of the expected heuristic running time of the AGS algorithm averaged over all problem instances, which was proved by Galbraith and Ruprai (Theorem 4, [24]). Assuming the walks are truly random, Theorem 5.2 gives the idealized asymptotic hazard rate of the AGS algorithm as $1 - F(r, x) = \lim_{n\to\infty} \mathbb{P}(T_N^{(x)}/\sqrt{N} > r)$, quantifying which problem instances are easier/harder. Figure 2(a) shows the asymptotic hazard rate for the AGS algorithm for various

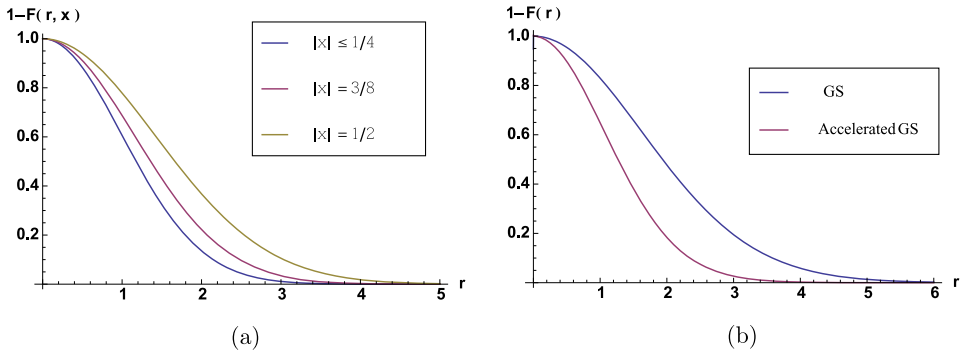

FIG. 2. (a) *Limiting idealized hazard rate of the AGS algorithm for various problem instances*, (b) *comparing limiting idealized hazard rates of GS and the AGS algorithms*.

values $x \in [-1/2, 1/2]$. It is observed that as x approaches $1/2$, the idealized running time of the AGS algorithm increases, which is expected as the intersection between the tame and wild sets decrease. Moreover, assuming that x uniformly distributed over $[-1/2, 1/2]$, the performance of the different variants of the GS algorithms can be compared using the limiting idealized hazard rates averaged over all problem instances

$$1 - F(r) = \int_{-1/2}^{1/2} \big(1 - F(r, x)\big)\, dx.$$

Figure 2(b) shows the limiting idealized hazard rates of the GS and AGS algorithms averaged over all problem instances. It shows that the limiting idealized running time of the GS algorithm stochastically dominates the AGS algorithm, that is, it is better than the GS algorithm not only in expectation but also for all values $r \geq 0$.

6. Collision times in sequential graph coloring using Stein's method. In this section, Stein's method for Poisson approximation is used to determine the limiting distributions of the collision times for the preferential attachment model and the infinite path. To this end, recall the following version of Stein's method based on dependency graph.

THEOREM 6.1 (Chatterjee et al. [13]). *Suppose $\{X_i\}_{i \in \mathscr{I}}$ is a finite collection of binary random variables with dependency graph $(\mathscr{I}, \mathscr{E})$, that is, $(X_i, X_j) \in \mathscr{E}$, whenever X_i, X_j are dependent. Let $W = \sum_{i \in \mathscr{I}} X_i$, $p_i = \mathbb{P}(X_i = 1)$, $p_{ij} = \mathbb{P}(X_i = X_j = 1)$, and $\lambda = \sum_{i \in \mathscr{I}} p_i$. Then*[4]

$$\big\| W - \mathrm{Pois}(\lambda) \big\| \leq \min\left\{1, \frac{1}{\lambda}\right\} \Bigg(\sum_{\substack{i \in \mathscr{I} \\ j \in N(i) \setminus \{i\}}} p_{ij} + \sum_{\substack{i \in \mathscr{I} \\ j \in N(i)}} p_i p_j \Bigg).$$

6.1. *Preferential attachment models*: *Proof of Theorem 1.4.* Recall the definition of the PA(m) model and the graph sequence $(G_m^t)_{t \geq 1}$ from Section 1.2. Denote by $S(G_m^t)$ the underlying simple graph associated with G_m^t. Let $\mathcal{G} = (G_m^t)_{t \geq 1}$ and consider the coloring scheme described in Section 1.2. Let $T_{N1}^{\mathrm{PA}(m)}$ be the first time there is a monochromatic edge in a sequential coloring of $\mathcal{G}$.

For $t \geq 1$ (possibly depending on N), let $X_{1,t}, X_{2,t}, \ldots$ be i.i.d. $\mathbb{P}_N$ [$X_{i,t}$ corresponds to the color of the vertex i in $S(G_m^t)$]. For $(i, j) \in E(S(G_m^t))$, define

$$Z_{(i,j)}^{(t)} = \mathbf{1}\{X_{i,t} = X_{j,t}\},$$

[4]Note that $||W - \mathrm{Pois}(\lambda)|| = \frac{1}{2} \sum_i |\mathbb{P}(W = i) - \frac{e^{-\lambda} \lambda^i}{i!}|$ is the total variation distance between W and the Poisson distribution with mean λ.

and $W_t = \sum_{(i,j)\in E(S(G_m^t))} Z_{(i,j)}^{(t)}$, which is the number of monochromatic edges in $S(G_m^t)$. Let $\lambda_t := \mathbb{E}(W_t) = |E(S(G_m^t))| \sum_i p_{Ni}^2 = (mt - o(t)) \sum_i p_{Ni}^2$.

For two distinct edges $e_1 = (i_1, j_1)$ and $e_2 = (i_2, j_2)$,

$$\mathbb{P}(Z_{e_1}^{(t)} = Z_{e_2}^{(t)} = 1) = \begin{cases} \Big(\sum_i p_{Ni}^2\Big)^2, & \text{if } e_1 \cap e_2 = \varnothing, \\ \sum_i p_{Ni}^3, & \text{otherwise,} \end{cases}$$

since the random variables $Z_{e_1}^{(t)}$ and $Z_{e_2}^{(t)}$ are independent whenever the edges $e_1 \cap e_2 = \varnothing$, the dependency graph associated with the random variables $\{Z_{(i,j)}^{(t)}\}_{(i,j)\in E(S(G_m^t))}$ can be constructed by putting an edge between $Z_{e_1}^{(t)}$ and $Z_{e_2}^{(t)}$ whenever e_1 and e_2 share a vertex. Then the error term in (6.1) becomes

$$\begin{aligned} \Big|\mathbb{P}(W_t = k|G_t) - e^{-\lambda_t}\frac{\lambda_t^k}{k!}\Big| &\le C_t 2|T(S(G_m^t))|\Big(\sum_i p_{Ni}^3 + \Big(\sum_i p_{Ni}^2\Big)^2\Big) \\ &\quad + 2C_t|E(S(G_m^t))|\Big(\sum_i p_{Ni}^2\Big)^2, \end{aligned} \tag{6.1}$$

where $C_t = \min(1, \lambda_t^{-1})$ and $T(S(G_m^t))$ is the number of 2-stars (the bipartite graph $K_{1,2}$) in $S(G_m^t)$.

If $t = t(N)$ is such that $\lim_{N\to\infty} t \sum_i p_{Ni}^2 = \lambda > 0$ for some $\lambda > 0$, then $\lambda_t \to m\lambda$, and $|E(S(G_m^t))|(\sum_i p_{Ni}^2)^2 \le p_{N1}\lambda_t \to 0$. Moreover, Bollobás [11], Theorem 16, shows that

$$(1-\varepsilon)\binom{m+1}{2} t\log t \le |S(G_m^t)| \le (1+\varepsilon)\binom{m+1}{2} t\log t,$$

with high probability as $t \to \infty$. Now, if $\lim_{N\to\infty} p_{N1}\log t = 0$, then with high probability

$$\lim_{N\to\infty}\lim_{t\to\infty} |T(S(G_m^t))| \sum_i p_{Ni}^3 \le \lambda(1+\varepsilon)\binom{m+1}{2} \lim_{N\to\infty}\lim_{t\to\infty} p_{N1}\log t = 0.$$

Therefore, the RHS goes to zero at $N \to \infty$ and by dominated convergence theorem, the number of monochromatic edges W_t converges in distribution to Pois(λ).

Thus, taking $t = \lfloor \frac{r}{s_N^2} \rfloor$, we get

$$\lim_{N\to\infty} \mathbb{P}(s_N^2 T_N^{\mathrm{PA}(m)} > r) = \lim_{N\to\infty} \mathbb{P}(W_{\lfloor \frac{r}{s_N^2} \rfloor} = 0) = e^{-mr},$$

completing the proof of Theorem 1.4.

6.2. *The infinite path.* Define $\mathcal{T}_{N,m}$ to be the first time there exists a monochromatic path of length m in a sequential coloring of the infinite path $\mathcal{Z}$ with the probability distribution $\mathbb{P}_N$. This problem can be re-formulated as follows: $\{X_i\}_{i\geq 1}$ be an i.i.d. $\mathbb{P}_N$ sequence, and

$$\mathcal{T}_{N,m} = \inf\{t \geq m : X_t = \cdots = X_{t-m+1}\}. \tag{6.2}$$

Similar to the proof of Theorem 1.4, using the Stein method based on dependency graph, the limiting distribution of $\mathcal{T}_{N,m}$ can be determined.

THEOREM 6.2. *Let* $s_{N,m} := (\sum_i p_{Ni}^m)^{\frac{1}{m}}$ *and suppose* $\lim_{N\to\infty} p_{N1} = 0$. *Then for* $r \geq 0$,

$$\lim_{N\to\infty} \mathbb{P}\big(s_{N,m}^m \mathcal{T}_{N,m} > r\big) = e^{-r}. \tag{6.3}$$

PROOF. For $t \geq m$, and $s \in [1, t-m+1]$, define

$$Z_{s,t} = \mathbf{1}\{X_s = X_{s+1} = \cdots = X_{s+m-1}\},$$

and $W_t = \sum_{s=1}^{t-m+1} Z_{s,t}$. Note that W_t counts the number of monochromatic paths with m vertices in the path spanned by the vertices $\{1, 2, \ldots, t\}$. Note that $\lambda_t := \mathbb{E}(W_t) = (t-m+1)\sum_i p_{Ni}^m$ and

$$\mathbb{P}(Z_{s_1,t} = Z_{s_2,t} = 1) = \begin{cases} \Big(\sum_i p_{Ni}^m\Big)^2, & \text{if } |s_2 - s_1| \geq m, \\ \sum_i p_{Ni}^{m-|s_2-s_1|}, & \text{otherwise.} \end{cases}$$

Since $Z_{s_1,t}$ and $Z_{s_2,t}$ are independent if $|s_2 - s_1| \geq m$, the dependency graph associated with the variables $\{Z_{s,t}\}_{s\geq 1}$ has an edge between $Z_{s_1,t}$ and $Z_{s_2,t}$ if $|s_2 - s_1| < m$. Then the error term in (6.1) becomes

$$\big\|W_i - \mathrm{Pois}(\lambda_i)\big\|_{\mathrm{TV}} \leq \frac{4m(t-m+1)(\sum_{b=0}^{m-1}\sum_i p_{Ni}^{m-b} + (\sum_i p_{Ni}^m)^2)}{(t-m+1)\sum_i p_{Ni}^m}.$$

The error term in the RHS goes to 0 if $t = t(N)$ is such that $\lambda_t = (t-m+1)\sum_i p_{Ni}^m \to \lambda > 0$, as $N \to \infty$.

For $r > 0$, let $t = \lceil \frac{r}{s_{N,m}^m} \rceil$. Then

$$\mathbb{P}\big(s_{N,m}^m \mathcal{T}_{N,m} > r\big) = \mathbb{P}(W_{\lfloor \frac{r}{s_{N,m}^m} \rfloor} = 0) \to e^{-r},$$

completing the proof of the result. □

APPENDIX: VERIFYING ABSOLUTE CONVERGENCE

For every $a \in [q]$, let $\underline{\psi}_a = (\psi_{1,a}, \psi_{2,a}, \ldots)$ and $\sum_i (\sum_{a=1}^q c_a \psi_{i,a})^2 < \infty$. Define

$$g(r, \underline{\psi}_1, \underline{\psi}_2, \ldots, \underline{\psi}_q) = \sum_i \left\{ -r \sum_{a=1}^q c_a \psi_{i,a} + \log\left(1 + \sum_{a=1}^q e^{r c_a \psi_{i,a}} - q\right)\right\}. \tag{A.1}$$

Also, let $Q_i(z) = \sum_{s=1}^\infty \gamma_{i,s} \frac{z^s}{s!}$ and $\gamma_{i,s} = \sum_{a=1}^q c_a^s \psi_{i,a}^s$. By standard rearrangement identities [20], for $s \geq 1$,

$$Q_i^s(z) = \sum_{\substack{j_1, j_2, \ldots, j_s \\ j_t \geq 1, \forall t \in [s]}} z^{\sum_{b=1}^s j_b} \left(\prod_{b=1}^s \frac{\sum_{a=1}^q c_a^{j_b} \psi_{i,a}^{j_b}}{j_b!} \right). \tag{A.2}$$

LEMMA A.3. *Let* $\sum_i (\sum_{a=1}^q c_a \psi_{i,a})^2 := C < \infty$, *and define* $\psi_{\max} = \max_{a \in [q]} \max_{i \in [n]} \psi_{i,a}$. *Then for* $r \in R = \{r \in \mathbb{R}^+ : \psi_{\max} r < \log(1 + 1/q)\}$,

$$\begin{aligned} &g(r, \underline{\psi}_1, \underline{\psi}_2, \ldots, \underline{\psi}_q) \\ &\quad = -\frac{r^2}{2} \sum_i \left(\left(\sum_a c_a \psi_{i,a} \right)^2 - \sum_{a=1}^q c_a^2 \psi_{i,a}^2 \right) \\ &\qquad + A(r, \underline{\psi}_1, \underline{\psi}_2, \ldots, \underline{\psi}_q) + B(r, \underline{\psi}_1, \underline{\psi}_2, \ldots, \underline{\psi}_q), \end{aligned} \tag{A.3}$$

where

$$\begin{aligned} &A(r, \underline{\psi}_1, \underline{\psi}_2, \ldots, \underline{\psi}_q) \\ &\quad = \sum_{j_1 \geq 3} \frac{r^{j_1}}{j_1!} \sum_i \sum_{a=1}^q c_a^{j_1} \psi_{i,a}^{j_1} - \frac{1}{2} \sum_{\substack{j_1 \geq 1, j_2 \geq 1, \\ j_1 + j_2 \geq 3}} r^{\sum_{b=1}^2 j_b} \sum_i \left(\prod_{b=1}^2 \frac{\sum_{a=1}^q c_a^{j_b} \psi_{i,a}^{j_b}}{j_b!} \right), \end{aligned}$$

and

$$B(r, \underline{\psi}_1, \underline{\psi}_2, \ldots, \underline{\psi}_q) = \sum_{s=3}^\infty \frac{(-1)^{s+1}}{s} Q_i^s(z).$$

Moreover, the series (A.3) *is absolutely convergent for* $r \in R$.

PROOF. For $r \in R$, $\psi_{i,a} < \frac{1}{r c_a} \log(1 + 1/q)$, for all $a \in [q]$ and $i \in \mathbb{N}$, and $|\sum_{a=1}^q e^{r c_a \psi_{i,a}} - q| < 1$. Thus, using the expansion of $\log(1 + z)$, for $|z| < 1$, and

the expansion of e^{-z}, for all $z \in \mathbb{R}$,

$$
\begin{aligned}
g(r, \underline{\psi}_1, \ldots, \underline{\psi}_q) &= \sum_i \left\{ -r \sum_{a=1}^{q} c_a \psi_{i,a} + \sum_{s=1}^{\infty} \frac{(-1)^{s+1}}{s} \left(\sum_{a=1}^{q} e^{r c_a \psi_{i,a}} - q \right)^s \right\} \\
&= \sum_i \left\{ -r \sum_{a=1}^{q} c_a \psi_{i,a} + \sum_{s=1}^{\infty} \frac{(-1)^{s+1}}{s} \left(\sum_{a=1}^{q} \sum_{t=1}^{\infty} \frac{r^t c_a^t \psi_{i,a}^t}{t!} \right)^s \right\} \qquad \text{(A.4)} \\
&= \sum_i \left\{ -r \sum_{a=1}^{q} c_a \psi_{i,a} + \sum_{s=1}^{\infty} \frac{(-1)^{s+1}}{s} Q_i(r)^s \right\}.
\end{aligned}
$$

Note that $Q_i(r) = r \sum_{a=1}^{q} c_a \psi_{i,a} + \sum_{j_1=2}^{\infty} \frac{r^{j_1}}{j_1!} (\sum_{a=1}^{q} c_a^{j_1} \psi_{i,a}^{j_1})$. Thus, (A.4) implies

$$
g(r, \underline{\psi}_1, \underline{\psi}_2, \ldots, \underline{\psi}_q) = \sum_i \left\{ \sum_{j_1=2}^{\infty} \frac{r^{j_1}}{j_1!} \left(\sum_{a=1}^{q} c_a^{j_1} \psi_{i,a}^{j_1} \right) + \sum_{s=2}^{\infty} \frac{(-1)^{s+1}}{s} Q_i^s(r) \right\}. \qquad \text{(A.5)}
$$

Define

$$
\mathscr{S} := \sum_i \left\{ \sum_{j_1=2}^{\infty} \frac{r^{j_1}}{j_1!} \left(\sum_{a=1}^{q} c_a^{j_1} \psi_{i,a}^{j_1} \right) + \sum_{s=2}^{\infty} \frac{Q_i^s(r)}{s} \right\}.
$$

To show the series $g(r, \underline{\psi}_1, \underline{\psi}_2, \ldots, \underline{\psi}_q)$ is absolutely convergent for $r \in R$, it suffices to show $\mathscr{S} < \infty$. To this end,

$$
\mathscr{S} \le \sum_{j_1=2}^{\infty} r^{j_1} \sum_i \sum_{a=1}^{q} c_a^{j_1} \psi_{i,a}^{j_1} + \sum_{s=2}^{\infty} \sum_i Q_i^s(r). \qquad \text{(A.6)}
$$

Note that, for $r \in R$,

$$
\begin{aligned}
\sum_{j_1=2}^{\infty} r^{j_1} \sum_i \sum_{a=1}^{q} c_a^{j_1} \psi_{i,a}^{j_1} &\le \sum_{j_1=2}^{\infty} r^{j_1} \sum_i \left(\sum_{a=1}^{q} c_a \psi_{i,a} \right)^{j_1} \le C \sum_{j_1=2}^{\infty} r^{j_1} \psi_{\max}^{j_1-2} \\
&\le \frac{C r^2}{(1 - r\psi_{\max})}.
\end{aligned}
$$

Similarly, recalling (A.2), it can shown that $\sum_{s=2}^{\infty} \sum_i Q_i^s(r) < \infty$, for $r \in R$. This implies that the series in the RHS of (A.6) is finite, and $g(r, \underline{\psi}_1, \ldots, \underline{\psi}_q)$ is absolutely convergent.

The result now follows by interchanging the order of the summations and rearranging the terms. □

Acknowledgement. The author is indebted to his advisor Persi Diaconis for proposing the problem to him and for his inspirational guidance. The author thanks Sourav Chatterjee, Shirshendu Ganguly, Prasad Tetali and Qingyuan Zhao for helpful discussions. The author also thanks an anonymous referee for carefully reading the manuscript and for providing many helpful comments which improved the quality and presentation of the paper.

REFERENCES

[1] ALDOUS, D. (1989). *Probability Approximations Via the Poisson Clumping Heuristic*. *Applied Mathematical Sciences* **77**. Springer, New York. MR0969362

[2] ARRATIA, R., GARIBALDI, S. and KILIAN, J. (2016). Asymptotic distribution for the birthday problem with multiple coincidences, via an embedding of the collision process. *Random Structures Algorithms* **48** 408–502.

[3] ATHREYA, K. B. and KARLIN, S. (1968). Embedding of urn schemes into continuous time Markov branching processes and related limit theorems. *Ann. Math. Statist.* **39** 1801–1817. MR0232455

[4] BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. MR2091634

[5] BARBOUR, A. D. and GNEDIN, A. V. (2009). Small counts in the infinite occupancy scheme. *Electron. J. Probab.* **14** 365–384. MR2480545

[6] BARBOUR, A. D., HOLST, L. and JANSON, S. (1992). *Poisson Approximation*. *Oxford Studies in Probability* **2**. Oxford Univ. Press, New York. MR1163825

[7] BATU, T., FORTNOW, L., RUBINFELD, R., SMITH, W. D. and WHITE, P. (2013). Testing closeness of discrete distributions. *J. ACM* **60** Art. 4, 25. MR3033221

[8] BHATTACHARYA, B. B., DIACONIS, P. and MUKHERJEE, S. (2014). Universal Poisson and Normal limit theorems in graph coloring problems with connections to extremal combinatorics. Available at arXiv:1310.2336.

[9] BILLINGSLEY, P. (1995). *Probability and Measure*, 3rd ed. Wiley, New York. MR1324786

[10] BOLLOBÁS, B. and RIORDAN, O. (2004). The diameter of a scale-free random graph. *Combinatorica* **24** 5–34. MR2057681

[11] BOLLOBÁS, B. and RIORDAN, O. M. (2003). Mathematical results on scale-free random graphs. In *Handbook of Graphs and Networks* 1–34. Wiley-VCH, Weinheim. MR2016117

[12] CAMARRI, M. and PITMAN, J. (2000). Limit distributions and random trees derived from the birthday problem with unequal probabilities. *Electron. J. Probab.* **5** no. 2, 18 pp. (electronic). MR1741774

[13] CHATTERJEE, S., DIACONIS, P. and MECKES, E. (2005). Exchangeable pairs and Poisson approximation. *Probab. Surv.* **2** 64–106. MR2121796

[14] DALEY, D. J. and VERE-JONES, D. (1988). *An Introduction to the Theory of Point Processes*. Springer, New York. MR0950166

[15] DASGUPTA, A. (2005). The matching, birthday and the strong birthday problem: A contemporary review. *J. Statist. Plann. Inference* **130** 377–389. MR2128015

[16] DIACONIS, P. and HOLMES, S. (2002). A Bayesian peek into Feller volume I. *Sankhyā Ser. A* **64** 820–841. MR1981513

[17] DIACONIS, P. and MOSTELLER, F. (1989). Methods for studying coincidences. *J. Amer. Statist. Assoc.* **84** 853–861. MR1134485

[18] DONG, F. M., KOH, K. M. and TEO, K. L. (2005). *Chromatic Polynomials and Chromaticity of Graphs*. World Scientific, Hackensack, NJ. MR2159409

[19] FADNAVIS, S. (2015). A note on the shameful conjecture. *European J. Combin.* **47** 115–122. MR3319082

[20] FLAJOLET, P. and SEDGEWICK, R. (2009). *Analytic Combinatorics*. Cambridge Univ. Press, Cambridge. MR2483235

[21] GALBRAITH, S. and RUPRAI, R. S. (2009). An improvement to the Gaudry-Schost algorithm for multidimensional discrete logarithm problems. In *Cryptography and Coding*. *Lecture Notes in Computer Science* **5921** 368–382. Springer, Berlin. MR2775634

[22] GALBRAITH, S. D. and HOLMES, M. (2012). A non-uniform birthday problem with applications to discrete logarithms. *Discrete Appl. Math.* **160** 1547–1560. MR2915391

[23] GALBRAITH, S. D., POLLARD, J. M. and RUPRAI, R. S. (2013). Computing discrete logarithms in an interval. *Math. Comp.* **82** 1181–1195. MR3008854

[24] GALBRAITH, S. D. and RUPRAI, R. S. (2010). Using equivalence classes to accelerate solving the discrete logarithm problem in a short interval. In *Public Key Cryptography—PKC* 2010. *Lecture Notes in Computer Science* **6056** 368–383. Springer, Berlin. MR2660753

[25] GAUDRY, P. and SCHOST, É. (2004). A low-memory parallel version of Matsuo, Chao, and Tsujii's algorithm. In *Algorithmic Number Theory*. *Lecture Notes in Computer Science* **3076** 208–222. Springer, Berlin. MR2137355

[26] GNEDIN, A., HANSEN, B. and PITMAN, J. (2007). Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws. *Probab. Surv.* **4** 146–171. MR2318403

[27] HOLST, L. (1986). On birthday, collectors', occupancy and other classical urn problems. *Internat. Statist. Rev.* **54** 15–27. MR0959649

[28] HOLST, L. (1995). The general birthday problem. In *Proceedings of the Sixth International Seminar on Random Graphs and Probabilistic Methods in Combinatorics and Computer Science*, *"Random Graphs '93"* (*Poznań*, 1993) **6** 201–208. MR1370955

[29] HOLST, L. (2001). Extreme value distributions for random coupon collector and birthday problems. *Extremes* **4** 129–145 (2002). MR1893869

[30] JENSEN, T. R. and TOFT, B. (1995). *Graph Coloring Problems*. Wiley, New York. MR1304254

[31] JOAG-DEV, K. and PROSCHAN, F. (1992). Birthday Problem with Unlike Probabilities. *Amer. Math. Monthly* **99** 10–12. MR1542029

[32] JOHNSON, N. L. and KOTZ, S. (1977). *Urn Models and Their Application*: *An Approach to Modern Discrete Probability Theory*. Wiley, New York. MR0488211

[33] KIM, J. H., MONTENEGRO, R., PERES, Y. and TETALI, P. (2010). A birthday paradox for Markov chains with an optimal bound for collision in the Pollard rho algorithm for discrete logarithm. *Ann. Appl. Probab.* **20** 495–521. MR2650040

[34] KINGMAN, J. F. C. (1993). *Poisson Processes*. *Oxford Studies in Probability* **3**. Oxford Univ. Press, New York. MR1207584

[35] MAHMOUD, H. M. (2009). *Pólya Urn Models*. CRC Press, Boca Raton, FL. MR2435823

[36] MONTENEGRO, R. and TETALI, P. (2009). How long does it take to catch a wild kangaroo? In *STOC'09—Proceedings of the* 2009 *ACM International Symposium on Theory of Computing* 553–559. ACM, New York. MR2780101

[37] NAKATA, T. (2008). A Poisson approximation for an occupancy problem with collisions. *J. Appl. Probab.* **45** 430–439. MR2426842

[38] NEAL, P. (2008). The generalised coupon collector problem. *J. Appl. Probab.* **45** 621–629. MR2455173

[39] PANINSKI, L. (2008). A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inform. Theory* **54** 4750–4755. MR2591136

[40] POLLARD, J. M. (1978). Monte Carlo methods for index computation (mod p). *Math. Comp.* **32** 918–924. MR0491431

[41] POLLARD, J. M. (2000). Kangaroos, Monopoly and discrete logarithms. *J. Cryptology* **13** 437–447. MR1788514

[42] SELIVANOV, B. I. (1995). On the waiting time in a scheme for the random allocation of colored particles. *Discrete Math. Appl.* **5** 73–82.

[43] STANLEY, R. P. (1995). A symmetric function generalization of the chromatic polynomial of a graph. *Adv. Math.* **111** 166–194. MR1317387

[44] WENDL, M. (2005). Probabilistic assessment of clone overlaps in DNA fingerprint mapping via a priori models. *J. Comput. Biol.* **12** 283–297.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: bhaswar@stanford.edu