GENERALIZATIONS OF POLYA'S URN PROBLEM

FAN CHUNG, SHIRIN HANDJANI, AND DOUG JUNGREIS

ABSTRACT. We consider generalizations of the classical Polya urn problem: Given finitely many bins each containing one ball, suppose that additional balls arrive one at a time. For each new ball, with probability p, create a new bin and place the ball in that bin; with probability $1-p$, place the ball in an existing bin, such that the probability the ball is placed in a bin is proportional to m^γ, where m is the number of balls in that bin. For $p = 0$, the number of bins is fixed and finite, and the behavior of the process depends on whether γ is greater than, equal to, or less than 1. We survey the known results and give new proofs for all three cases. We then consider the case $p > 0$. When $\gamma = 1$, this is equivalent to the so-called *preferential attachment scheme* which leads to power law distribution for bin sizes. When $\gamma > 1$, we prove that a single bin dominates, *i.e.,* as the number of balls goes to infinity, the probability converges to 1 that any new ball either goes into that bin or creates a new bin. When $p > 0$ and $\gamma < 1$, we show that under the assumption that certain limits exist, the fraction of bins having m balls shrinks exponentially as a function of m. We then discuss further generalizations and pose several open problems.

Research supported in part by NSF Grants DMS 0100472 and ITR 0205061.
This paper is to appear in *Annals of Combinatorics*.

1. Introduction

We consider the following process involving balls and bins:

> For fixed parameters, $\gamma \in \mathbf{R}$, $0 \leq p < 1$ and a positive integer $k > 1$, begin with k bins, each containing one ball and then introduce balls one at a time. For each new ball, with probability p, create a new bin and place the ball in that bin; with probability $1 - p$, place the ball in an existing bin, such that the probability the ball is placed in a bin is proportional to m^γ, where m is the number of balls in that bin.

A well-known result of this type is Polya's urn problem (see [8]), which is just the above process in the case of $p = 0$ and $\gamma = 1$. For general γ, if p is 0, then we never create new bins, and we refer to the process as a *finite Polya process* with exponent γ. If $p \neq 0$, then we refer to it as an *infinite Polya process* with exponent γ. If $\gamma > 1$, the process is often considered as having *positive feedback* by economists modeling the tendency toward monopoly[2, 3, 12]. Similarly, for $\gamma < 1$, the process is regarded as having *negative feedback* in various situations concerning decreasing advantage in competition [12].

Many results in the literature on the finite Polya process (except for the classical Polya's urn problem) are either non-rigorous or folklore. Recently, Drinea et al. [7] analyzed the process for two bins with feedback. Spencer and Wormald [13] examined the case of many bins. The behavior of the finite Polya process can be broken into three cases depending on the value of γ. For $\gamma < 1$, the bins all grow at roughly the same rate. For $\gamma > 1$, one bin dominates, *i.e.,* the probability goes to 1 that any new ball goes into that bin. For $\gamma = 1$, the fraction of balls going into each bin converges, though the limit is uniformly distributed in a certain simplex. For completeness, we will give new short proofs for all three cases in Section 2.

The infinite Polya process is motivated by a variation of a web tree-graph model proposed by Drinea et al. [6] and by Kaprivsky and Redner [9]. The *preferential attachment* scheme (see Barabási et al [1, 4]) was further generalized so that the probability of a new node linking to an existing node with in-degree m is proportional to m^γ for some γ. This web tree-graph model is essentially equivalent to the infinite Polya process with $p = 1/2$ and exponent γ. In this paper, we examine the infinite Polya process for all ranges of γ and p. In Section 3, we prove that for $\gamma > 1$, a single bin dominates, *i.e.,* as the number of balls goes to infinity, the probability goes to 1 that any new ball either goes into that bin or creates a new bin. In addition, for any integer m satisfying $m < (m-1)\gamma$, only finitely many bins ever reach size m; whereas, for m satisfying $m > (m-1)\gamma$, the expected number of bins of size m at time t is of order $t^{m-(m-1)\gamma}$. In Section 4, we consider the infinite Polya process with exponent $\gamma \leq 1$. For $\gamma = 1$, the process generates a power law distribution so that the fraction of bins having m balls is asymptotic to $cm^{-\beta}$, where $\beta = 1 + 1/(1-p)$ and c is a constant. For $\gamma < 1$, we conjecture but do not prove that the limit of

the fraction of bins having m balls exists. Under certain assumptions concerning the convergence of these limits, we derive that the fraction of bins having m balls shrinks exponentially as a function of m. In the last section, we summarize the current state of the affair and mention a number of open problems.

2. Finite Case

In this section, we consider the finite Polya process. Throughout, we measure time t by the total number of balls in bins.

Theorem 2.1. *Consider a finite Polya process with exponent $\gamma = 1$ and k bins, and let x_i^t denote the fraction of balls in the i^{th} bin at time t. Then almost surely (a.s.) for each i, the limit $X_i = \lim_{t\to\infty} x_i^t$ exists. Furthermore these limits are distributed uniformly on the simplex $\{(X_1, \ldots, X_k) : X_i > 0, X_1 + \cdots + X_k = 1\}$.*

Proof. Given n objects labeled $1, \ldots, n$, we view a permutation $\pi \in S_n$ as a listing of the objects in the order $\pi(1), \ldots, \pi(n)$. If starting with a permutation $\pi \in S_n$ and removing the object labeled n leaves us with the permutation $\rho \in S_{n-1}$, then we say that π *contains* ρ. Let P denote the set of infinite sequences $(\pi_1, \pi_2, \ldots)$ such that $\pi_i \in S_i$ and π_{i+1} contains π_i, for each i. Let the uniform probability measure on P refer to the measure for which each of the $i!$ choices for π_i is equally likely. We can sample uniformly from P as follows: begin with π_1 (*i.e.*, the object labeled 1), and proceed recursively; having placed the first i objects, place the object labeled $i+1$ in any of its $i+1$ possible positions with equal probability.

An element of S_n represents a configuration of $n+1$ balls in k bins as follows: view the objects labeled $1, \ldots, (k-1)$ as division points, which subdivide the list of objects into k sublists; view each sublist as a bin, and assume that each bin initially contains one ball; view each of the objects $k, \ldots, n$ in any sublist as an additional ball placed in that bin. If we sample uniformly from P as described above, and let the resulting permutation represent a configuration of balls and bins, then each time we place a ball, the probability it goes into any bin is proportional to the number of balls currently in that bin. Therefore, this is equivalent to the finite Polya process.

An alternative way to sample uniformly from P is to select an i.i.d. sequence of numbers (y_i) uniformly from the interval $[0, 1]$, and to define π_n to be the permutation of $1, \ldots, n$ produced by sorting according to the values y_i. The numbers $y_1, \ldots, y_{k-1}$ subdivide the interval into k subintervals, which represent the k bins. If we denote the lengths of these subintervals by $X_1, \ldots, X_k$, then the X_i are distributed as in the statement of the theorem. The fraction of balls that will go into the i^{th} bin is the fraction of y's in the i^{th} subinterval, which converges to X_i. □

As a consequence, we can immediately compute the distribution of X_1, and hence any X_i. Let f denote the density function for X_1, and define $v(x)$ to be the volume of the $(k-2)$-dimensional simplex $\{(X_2, \ldots, X_k) : X_i > 0, X_2 + \cdots + X_k = 1 - x\}$.

Then

$$f(x) = \frac{v(x)}{\int_0^1 v(y)\,dy} = \frac{(1-x)^{k-2}}{\int_0^1 (1-y)^{k-2}\,dy} = (k-1)(1-x)^{k-2}.$$

The probability that the interval X_1 has length at least x/k (*i.e.*, the first bin has at least x times the average number of balls) is

$$\int_{x/k}^1 f(y)\,dy = (1-x/k)^{k-1},$$

which is approximately e^{-x} when k is much larger than x.

Lemma 2.1. *Consider two finite Polya processes, both with k bins, and with exponents γ and 1 respectively, where $\gamma > 1$. Let n_i^t denote the total number of balls in the i largest bins at time t for the first process, and define m_i^t similarly for the second process. Then it is possible to couple the two processes so that $n_i^t \geq m_i^t$ for all i and t.*

Proof. We couple the two processes as follows: Let p_i^t (resp. q_i^t) denote the probability a ball goes into any of the i largest bins for the first (resp. second) process. (In case there is a tie for the i^{th} largest bin, arbitrarily call one of the tied bins the i^{th} largest.) To place a ball at time t for the two processes, we select a random number y uniformly from $[0,1]$ and for the first (resp. second) process, we place a ball in the i^{th} largest bin for the smallest i such that $y < p_i^t$ (resp. $y < q_i^t$).

We now show that with this coupling the first process always has at least as many total balls in the i largest bins, for all i. Initially, the processes have the same number of balls, so we proceed by induction. Suppose it is true for all i at time t. We will prove it for an arbitrary i at time $t+1$. For the first process at time t, denote the bin sizes $b_1, \ldots, b_k$ from largest to smallest. For the second process, denote them $c_1, \ldots, c_k$. If $b_1 + \cdots + b_i$ is strictly greater than $c_1 + \cdots + c_i$, then we are done, so suppose these are equal. Then we have

$$\begin{aligned}
p_i^t &= \frac{\sum_{j=1}^i b_j^\gamma}{\sum_{j=1}^k b_j^\gamma} \geq \frac{\sum_{j=1}^i b_j^\gamma}{\sum_{j=1}^i b_j^\gamma + \sum_{j=i+1}^k b_j b_i^{\gamma-1}} \\
&= 1 - \frac{\sum_{j=i+1}^k b_j b_i^{\gamma-1}}{\sum_{j=1}^i b_j^\gamma + \sum_{j=i+1}^k b_j b_i^{\gamma-1}} \\
&\geq 1 - \frac{\sum_{j=i+1}^k b_j b_i^{\gamma-1}}{\sum_{j=1}^i b_j b_i^{\gamma-1} + \sum_{j=i+1}^k b_j b_i^{\gamma-1}} \\
&= 1 - \frac{\sum_{j=i+1}^k b_j}{\sum_{j=1}^k b_j} = 1 - \frac{\sum_{j=i+1}^k c_j}{\sum_{j=1}^k c_j} = q_i^t.
\end{aligned}$$

Thus $p_i^t \geq q_i^t$, which means that if we place a ball in one of the i largest bins for the second process, then we will do so for the first process also. This would complete the

induction, except for the case of ties: suppose there is a tie for the i^{th} largest bin, *i.e.*, $c_i = c_{i+1} = \cdots = c_j$. This is relevant because adding a ball to any of these bins will make it the i^{th} largest, thereby adding one to the size of the i largest bins. Since $b_1 + \cdots + b_i = c_1 + \cdots + c_i$, and $b_1 + \cdots + b_{i-1} \geq c_1 + \cdots + c_{i-1}$, we have $b_i \leq c_i$. But we also have $b_1 + \cdots + b_j \geq c_1 + \cdots + c_j$, so $b_i = \cdots = b_j = c_i = \cdots = c_j$. Then if either process gets a ball in one of its j largest bins, it will increase by 1 the total size of its i largest bins. We can then prove the induction step for i by instead proving the induction step for j. But since $c_j \neq c_{j+1}$, we can ignore the situation where there are ties. □

Lemma 2.2. *Consider two finite Polya processes, both with k bins, and with exponents γ and 1 respectively, where $\gamma < 1$. Let n_i^t denote the total number of balls in the i largest bins at time t for the first process, and define m_i^t similarly for the second process. Then it is possible to couple the two processes so that $n_i^t \leq m_i^t$ for all i and t.*

Proof. The proof is exactly the same as the previous lemma, except the inequalities are reversed. □

Lemma 2.3. *Given a finite or infinite Polya process with exponent γ and an arbitrary finite initial configuration (*i.e.*, finitely many balls arranged in finitely many bins), suppose we restrict attention to any particular subset of the bins and ignore any balls that are placed in the other bins. Then the process behaves exactly like a finite Polya process with exponent γ on this subset of bins, though the process may terminate after finitely many balls.*

Proof. If we condition on a ball being placed in this subset of bins, then the probability it is placed in each bin of the subset is still proportional to m^γ, where m is the number of balls in that bin. □

Lemma 2.4. *Consider a finite Polya process with exponent $\gamma > 1$ and only two bins. Let x_i^t denote the fraction of balls in the i^{th} bin at time t. Then a.s. the limit $X_i = \lim_{t\to\infty} x_i^t$ exists for each i; one of these limits is 1 and the other is 0.*

Proof. Compare this process to a second two-bin finite process with exponent 1. By Theorem 2.1, for the second process, x_i^t converges for each i, and with probability 1, the limits are not equal. Thus for any $\epsilon_1 > 0$, there is some $\alpha > 1$ such that with probability at least $1 - \epsilon_1$, the larger bin is eventually always greater than α times as large as the smaller bin. Then by Lemma 2.1, for the first process, with probability at least $1 - \epsilon_1$ the larger bin is eventually always at least α times as large as the smaller bin. Then for each subsequent ball, the probability it goes into the larger bin is at least α^γ times that for the other bin. The larger bin is therefore eventually always at least $\alpha^{\gamma-\epsilon_2}$ times as large as the other bin, for any $\epsilon_2 > 0$. Repeating this argument,

with probability at least $1-\epsilon_1$, the ratio of the larger bin size to the smaller goes to infinity. □

Lemma 2.5. *Consider a finite Polya process with exponent $\gamma < 1$ and only two bins. Let x_i^t denote the fraction of balls in the i^{th} bin at time t. Then a.s. the limit $X_i = \lim_{t\to\infty} x_i^t$ exists for each i, and both limits are $1/2$.*

Proof. If $\gamma < 0$, then each ball is more likely to go in the smaller bin (until it catches up), so the smaller bin is eventually always at least $1-\epsilon_2$ times as large as the larger bin, for any ϵ_2. Assume then that $\gamma \geq 0$. Compare this process to a second two-bin finite process with exponent 1. By Theorem 2.1, for the second process, x_i^t converges for each i, and with probability 1, neither limit is 0. Thus for any $\epsilon_1 > 0$, there is some $\alpha > 0$ such that with probability at least $1-\epsilon_1$, the smaller bin is eventually always greater than α times as large as the larger bin. Then by Lemma 2.2, for the first process, with probability at least $1-\epsilon_1$ the smaller bin is eventually always at least α times as large as the larger. Then for each subsequent ball, the probability it is placed in the smaller bin is at least α^γ times that for the larger bin, and the smaller bin is therefore eventually always at least $\alpha^{\gamma+\epsilon_2}$ times as large as the larger bin, for any $\epsilon_2 > 0$. Repeating this argument, with probability at least $1-\epsilon_1$, the ratio of the smaller bin size to the larger converges to 1. □

Theorem 2.2. *Consider a finite k-bin Polya process with exponent γ, and let x_i^t denote the fraction of balls in bin i at time t. Then a.s. the limit $X_i = \lim_{t\to\infty} x_i^t$ exists for each i. If $\gamma > 1$, then $X_i = 1$ for one bin, and $X_i = 0$ for the other bins. If $\gamma < 1$, then $X_i = \frac{1}{k}$ for all bins.*

Proof. Consider first the case $\gamma > 1$. Lemmas 2.3 and 2.4 imply that of any two bins, at least one must have the property that the probability the t^{th} ball is placed in that bin converges to 0 as t goes to infinity. Thus X_i exists and is 0 for all but one bin. Then X_i exists and is 1 for the remaining bin.

Now suppose $\gamma < 1$. Combining Lemmas 2.3 and 2.5, since some bin must grow arbitrarily large, they all must, and the ratios of their sizes must all converge to 1. Therefore all of the X_i exist, and they are all $\frac{1}{k}$. □

3. Infinite Case with Positive Feedback

In this section, we consider infinite Polya processes with exponent $\gamma > 1$. Again, we measure time t by the total number of balls in bins. Let $\max_t$ denote the maximum number of balls in any bin at time t.

Lemma 3.1. *For any infinite Polya process with $\gamma > 1$, a.s. $\lim_{t\to\infty} \max_t = \infty$.*

Proof. At time t, there are at most t bins, and the largest bin is at least as likely as any other bin to get the next ball, so the largest bin gets another ball with probability at least $\frac{1-p}{t}$. Since $\sum_{t=1}^{\infty} \frac{1-p}{t} = \infty$, the result follows. □

Lemma 3.2. *Given an infinite Polya process with exponent $\gamma > 1$, fix any $\epsilon > 0$. For sufficiently large t, the largest bin gets the $(t+1)^{\text{st}}$ ball with probability at least $\frac{(1-\epsilon)\max_t}{t}$.*

Proof. Select $\delta > 0$ such that $\frac{1-p}{1-p(1-\delta)^2} > 1-\epsilon$. Each ball creates a new bin with probability p, so $1/t$ times the number of bins at time t converges to p, and the average bin size converges to $1/p$. If we select $N > \frac{1}{p\delta}$, then eventually the fraction of bins that have size less than N is at least $(1-\delta)$. Thus eventually the fraction of balls in bins of size less than N is at least $(1-\delta)p$. Now select M such that $M^{\gamma-1} > N^{\gamma-1}/\delta$. Denote the bin sizes at some time t by $b_1, \ldots, b_k$, and suppose that the bins of size less than N are $b_1, \ldots, b_j$. By the previous lemma, if t is sufficiently large, then the largest bin has size greater than M, and the probability that the largest bin gets the $(t+1)^{\text{st}}$ ball is:

$$\begin{aligned}
\frac{(1-p)\max_t^\gamma}{\sum_{i=1}^k b_i^\gamma} &\geq \frac{(1-p)\max_t \max_t^{\gamma-1}}{\sum_{i=1}^j b_i N^{\gamma-1} + \sum_{i=j+1}^k b_i \max_t^{\gamma-1}} \\
&\geq \frac{(1-p)\max_t}{\sum_{i=1}^j b_i \delta + \sum_{i=j+1}^k b_i} \\
&\geq \frac{(1-p)\max_t}{(1-\delta)p\delta t + (1-(1-\delta)p)t} \\
&= \frac{\max_t}{t} \cdot \frac{1-p}{1-p(1-\delta)^2} \\
&> \frac{(1-\epsilon)\max_t}{t}.
\end{aligned}$$

□

Lemma 3.3. *Suppose we independently place t balls in bins such that each ball has probability at least m/t of landing in the first bin. Then the probability that the first bin receives fewer than $m-c$ balls is less than $me^{-c^2/(2m)}$.*

Proof. It will suffice to prove this assuming that each ball has probability exactly m/t. Let p_k denote the probability the first bin receives exactly k balls. Then

$$\begin{aligned}
p_{m-c-1} &< \frac{p_{m-c-1}}{p_m} = \frac{\binom{t}{m-c-1}\left(\frac{m}{t}\right)^{m-c-1}\left(\frac{t-m}{t}\right)^{t-m+c+1}}{\binom{t}{m}\left(\frac{m}{t}\right)^{m}\left(\frac{t-m}{t}\right)^{t-m}} \\
&= \prod_{i=1}^{c+1} \frac{(m-i+1)(t-m)}{(t-m+i)m} \leq \prod_{i=1}^{c+1} \frac{(m-i+1)}{m} \leq \prod_{i=1}^{c} e^{-i/m} \leq e^{-c^2/(2m)}
\end{aligned}$$

Then $\sum_{i=0}^{m-c-1} p_i < m p_{m-c-1} < m e^{-c^2/(2m)}$. □

Lemma 3.4. *Given an infinite Polya process with exponent $\gamma > 1$, for any $\epsilon > 0$, a.s. we eventually have* $\max_t > t^{1-\epsilon}$.

Proof. Select three numbers $\delta_1, \delta_2, \delta_3 > 0$ sufficiently small that they satisfy the following two conditions: $\frac{(1-\delta_2)(1-\delta_3)}{(1+\delta_1)} > 1 - \epsilon/2$ and $1 + \delta_1(1 - \epsilon/2) > (1 + \delta_1)^{(1-\epsilon)}$. Let ρ denote the quantity $\frac{(1-\delta_2)\delta_1}{(1+\delta_1)}$. Combining the two conditions above, we have: $1 + (1 - \delta_3)\rho > (1 + \delta_1)^{(1-\epsilon)}$.

Let t_i denote the first time for which the largest bin contains i balls. We consider the interval of time from t_i to $(1+\delta_1)t_i$, which we refer to as the i^{th} time interval. We will show that during the i^{th} time interval, the largest bin is likely to grow by more than a factor of $(1 + \delta_1)^{(1-\epsilon)}$.

During the i^{th} time interval, time is never more than $(1 + \delta_1)t_i$, and the size of the largest bin is never less than i, so Lemma 3.2 implies that if t_i is sufficiently large, then each ball during this time interval goes into the largest bin with probability greater than $\frac{(1-\delta_2)i}{(1+\delta_1)t_i}$. Then the expected number of balls that go into the largest bin during this time interval is at least $\frac{(1-\delta_2)\delta_1 i}{(1+\delta_1)} = \rho i$, so by Lemma 3.3, the probability that the largest bin receives fewer than $(1 - \delta_3)\rho i$ balls is at most $\rho i e^{-(\delta_3\rho i)^2/(2\rho i)} = \rho i e^{-\delta_3^2\rho i/2}$. The sum from i equals zero to infinity of this expression is finite, so there can be at most finitely many i for which the largest bin fails to receive $(1 - \delta_3)\rho i$ balls during the i^{th} time interval. Then eventually, whenever time grows by a factor of $1 + \delta_1$, the largest bin grows by a factor of $1 + (1 - \delta_3)\rho > (1 + \delta_1)^{(1-\epsilon)}$, and so eventually $\max_t > t^{1-\epsilon}$. □

Theorem 3.1. *Given an infinite Polya process with exponent $\gamma > 1$, a.s. there is one bin such that the probability a ball is either placed in that bin or creates a new bin converges to 1. Also, for any $k \in \mathbf{Z}^+$ such that $k < (k - 1)\gamma$, only finitely many bins ever reach size k.*

Proof. Fix any $k \in \mathbf{Z}^+$ such that $k < (k - 1)\gamma$, and select $\epsilon > 0$ such that $k < (k - 1)\gamma(1 - \epsilon)$. To simplify notation, we define $\gamma' = \gamma(1 - \epsilon)$. By Lemma 3.4, eventually (for large time t) the largest bin has size at least $t^{1-\epsilon}$. Suppose this is true for all times $t > N$. Then the probability that a ball goes into any particular bin of size i at time $t > N$ is bounded by $\frac{i^\gamma}{t^{(1-\epsilon)\gamma}} = \frac{i^\gamma}{t^{\gamma'}}$. Now suppose a bin is created at time $t_1 > N$. The probability that it then receives balls at times $t_2 < \cdots < t_k$ (but at no other times before t_k) is at most $\prod_{i=1}^{k-1} \frac{i^\gamma}{t_{i+1}^{\gamma'}}$. We can now sum this over values of

$t_2, \ldots, t_k$ to get a bound on the probability that this bin ever receives k balls:

$$\begin{aligned}\sum_{\{(t_2,\ldots,t_k):t_1<\cdots<t_k\}} \prod_{i=1}^{k-1} \frac{i^\gamma}{t_{i+1}^{\gamma'}} &< ((k-1)\,!)^\gamma \int_{t_1}^\infty \cdots \int_{t_{k-1}}^\infty (t_2 \ldots t_k)^{-\gamma'}\, dt_k \ldots dt_2 \\ &= \frac{((k-1)\,!)^\gamma}{(\gamma'-1)(2\gamma'-2)\ldots((k-1)\gamma'-(k-1))} t_1^{-(k-1)\gamma'+(k-1)} \\ &= \frac{((k-1)\,!)^{\gamma-1}}{(\gamma'-1)^{k-1}} t_1^{-(k-1)\gamma'+(k-1)}.\end{aligned}$$

If we then integrate this expression from N to ∞, we get a bound on the expected number of bins that are created after time N and receive at least k balls:

$$\frac{((k-1)\,!)^{\gamma-1}}{(\gamma'-1)^{k-1}((k-1)\gamma'-k)} N^{-(k-1)\gamma'+k}.$$

This expected value is finite, so only finitely many bins can ever receive k balls.

By Lemma 3.4, for large t, the largest bin has at least $t^{1-\epsilon}$ balls; and since there are at most t bins, the probability that a ball goes into an existing bin of size less than k at time t is less than

$$\frac{tk^\gamma}{t^{(1-\epsilon)\gamma}},$$

which converges to 0 as t gets large. Thus the probability converges to 1 that if a ball is placed in an existing bin, then it is placed in one of the finitely many bins with at least k balls. Combining Lemma 2.3 and Theorem 2.2, the fraction of balls that are placed in all but one of these bins converges to 0, and so the probability that the t^{th} ball is placed in all but one of these bins converges to 0. Thus the probability converges to 1 that the t^{th} ball either is placed in the remaining bin or creates a new bin. □

Corollary 3.1. *Given an infinite Polya process with exponent $\gamma > 1$, for any $k \in \mathbf{Z}$ such that $k > (k-1)\gamma$, the expected number of bins of size k at time t is of order $t^{k-(k-1)\gamma}$* (i.e., *is eventually bounded above and below by expressions of the form $\alpha t^{k-(k-1)\gamma}$).*

Proof. If the bin sizes at some time t are $b_1, \ldots, b_m$, then $\sum_{i=1}^m b_i^\gamma \le t^\gamma$. Also by Theorem 3.1, for any $\epsilon > 0$, if t is large enough, then the size of the largest bin is at least $(1-p-\epsilon)t$, where $1-p$ is the probability a ball is placed in an existing bin. Therefore, $\sum_{i=1}^m b_i^\gamma \ge (1-p-\epsilon)^\gamma t^\gamma$. Thus for large t, the probability that a ball goes into any particular bin of size i is between $\frac{(1-p)i^\gamma}{t^\gamma}$ and $\frac{(1-p)i^\gamma}{(1-p-\epsilon)^\gamma t^\gamma}$; in other words, the probability is within a constant factor of $t^{-\gamma}$.

Now suppose a bin is created at some time t_1 where t_1 is sufficiently large that the above bounds hold. The probability that the bin receives additional balls at times $t_2 < t_3 < \cdots < t_k$ but at no other times is within a constant factor of $\prod_{i=1}^{k-1} \frac{i^\gamma}{t_{i+1}^\gamma}$.

(Note that the requirement that the bin does not receive a ball at any other time only introduces a constant factor, since $t^{-\gamma}$ is summable.) Summing over values for $t_2, \ldots, t_k$, we get:

$$\sum_{\{(t_2,\ldots,t_k):t_1<\cdots<t_k<t\}} \prod_{i=1}^{k-1} \frac{i^\gamma}{t_{i+1}^\gamma},$$

which for large t_1 is approximately equal to

$$((k-1)!)^\gamma \int_{t_1}^{t} \ldots \int_{t_{k-1}}^{t} (t_2 \ldots t_k)^{-\gamma}\, dt_k \ldots dt_2.$$

At each time t_1, a new bin is created with probability $1-p$, so we can now sum over large $t_1 < t$ to see that the expected number of bins of size k at time t is (up to a constant factor) $t^{-(k-1)\gamma+k}$. □

4. Infinite Case with $\gamma \leq 1$

We now consider the case $p > 0$ and $\gamma \leq 1$. Let $f_{i,t}$ denote the fraction of bins at time t that contain exactly i balls. It seems clear that $\lim_{t\to\infty} f_{i,t}$ exists for each i; however, we do not know how to prove this (except in the case $\gamma = 1$ where it is straightforward to induct on i). Since we cannot prove the limits exist, in this section we simply derive the behavior of the process under the following assumptions:

Assumptions (∗):

(1) For each i, there exists $f_i \in \mathbf{R}^+$ such that a.s. $\lim_{t\to\infty} f_{i,t}$ exists and is equal to f_i.
(2) A.s., $\lim_{t\to\infty} \sum_{j=1}^{\infty} f_{j,t} j^\gamma$ exists, is finite, and is equal to $\sum_{j=1}^{\infty} f_j j^\gamma$.

We remark that assumption (2) of (∗) does not hold for the case of $\gamma > 1$. In fact, by Theorem 3.1, when $\gamma > 1$, $\lim_{t\to\infty} \sum_{j=1}^{\infty} f_{j,t} j^\gamma = \infty$; whereas $f_j = 0$ for $j > 1$, and $f_1 = 1$, so $\sum_{j=1}^{\infty} f_j j^\gamma = 1$.

Theorem 4.1. *For an infinite Polya process with exponent $\gamma \leq 1$, suppose assumptions (∗) hold. Then there exists a constant $K > 0$ (depending on p and γ) such that for $i \geq 2$, a.s.*

$$f_i = \left(\frac{(i-1)^\gamma}{K + i^\gamma}\right) f_{i-1}. \tag{1}$$

Proof. Let $p_{i,t}$ denote the probability that the ball at time t is placed in a bin of size i, with the convention that $p_{0,t} = p$. Then for $i > 0$,

$$p_{i,t} = \frac{(1-p) f_{i,t} i^\gamma}{\sum_{j=1}^{\infty} f_{j,t} j^\gamma}.$$

Let $E_{i,t}$ denote the expected change in the number of bins of size i at time t, so $E_{i,t} = p_{i-1,t} - p_{i,t}$. By our assumptions, $p_{i,t}$ has a limit as t goes to infinity, so $E_{i,t}$

does as well. Call these limits p_i and E_i. The E_i must be in the same proportion as the f_i, that is, $\frac{E_i}{f_i} = \frac{E_j}{f_j}$, or else the fractions of bins of sizes i and j would eventually change. Defining C to be the constant value of $\frac{E_i}{f_i}$, we get:

$$Cf_i = E_i = p_{i-1} - p_i = \frac{(1-p)(f_{i-1}(i-1)^\gamma - f_i i^\gamma)}{\sum_{j=1}^\infty f_j j^\gamma},$$

for $i \geq 2$. Defining a new constant $K = C\sum_{j=1}^\infty f_j j^\gamma/(1-p)$, we get the desired expression. □

In the next theorem, we use the recurrence (1) to estimate the values f_i for large i, assuming that $(*)$ holds. We use the notation $f_i \propto g(i)$ to mean that $f_i = c(1 + o(1))g(i)$ for some constant c.

Theorem 4.2. *For an infinite Polya process with exponent $\gamma \leq 1$, suppose that the assumptions $(*)$ hold. Then the limit f_i of the fraction of bins with i balls a.s. satisfies the following:*

$$f_i \quad \propto \quad \begin{cases} i^{-(1+1/(1-p))} & \text{if } \gamma = 1, \\ i^{-\gamma} e^{-Ki^{1-\gamma}/(1-\gamma)} & \text{if } 0 < \gamma < 1, \\ (K+1)^{-i} & \text{if } \gamma = 0, \\ O\left(\frac{((i-1)!)^\gamma}{K^i}\right) & \text{if } \gamma < 0. \end{cases}$$

Proof. Define E_i, $E_{i,t}$, p_i, $p_{i,t}$, C, and K as in the previous proof. We first show that $\sum_{i=1}^\infty f_i = 1$. At any time t, the sum over all i of the fraction of bins that have size i must be 1, so $\sum_{i=1}^\infty f_{i,t} = 1$. It clearly follows that $\sum_{i=1}^\infty f_i \leq 1$, so suppose that $\sum_{i=1}^\infty f_i < 1 - \frac{1}{n}$ for some $n \in \mathbf{Z}^+$. Then for sufficiently large t, $\sum_{i=1}^{4n/p} f_{i,t} < 1 - \frac{1}{2n}$, which means that $\sum_{i=4n/p}^\infty f_{i,t} > \frac{1}{2n}$, and so the average bin size at time t is $\sum_{i=1}^\infty f_{i,t} i > \frac{4n}{p}\frac{1}{2n} = \frac{2}{p}$. But this contradicts the fact that the average bin size converges to $\frac{1}{p}$, so we must have $\sum_{i=1}^\infty f_i = 1$.

We next show that $\sum_{i=1}^\infty E_i = p$. At any time t, the expected change in the number of bins of size $\leq n$ is $\sum_{i=1}^n E_{i,t} = p - p_{n,t}$, so $\sum_{i=1}^n E_i = p - p_n$. Since $\sum_{n=1}^\infty p_n$ is bounded, $\lim_{n\to\infty} p_n = 0$, so $\sum_{i=1}^\infty E_i = \lim_{n\to\infty} p - p_n = p$.

Thus $C = \frac{E_i}{f_i} = p$, for all $\gamma \leq 1$, so $K = \frac{p}{1-p}\sum_{i=1}^\infty f_i i^\gamma$.

Now consider the case $\gamma = 1$, where we can evaluate the constant K explicitly. The average bin size converges to $\sum_{i=1}^\infty f_i i = \frac{1}{p}$ (using Assumption (2) of $(*)$), so $K = \frac{1}{1-p}$. We now use Theorem 4.1 to compute the limiting behavior of f_i as i gets large:

$$f_i \propto \prod_{j=2}^{i} \frac{j-1}{j+1/(1-p)} \propto \frac{\Gamma(i)}{\Gamma(i+1+1/(1-p))} \propto i^{-(1+\frac{1}{1-p})}$$

where Γ is the well-known Gamma function. Thus, the bin sizes in this case obey a power-law distribution with power law exponent $1 + 1/(1-p)$.

<table>
<tr><td></td><td>Finite Polya process
$p = 0$</td><td colspan="2">Infinite Polya process
$0 < p < 1$</td></tr>
<tr><td>$\gamma > 1$</td><td>one bin dominates</td><td colspan="2">one bin dominates</td></tr>
<tr><td>$\gamma = 1$</td><td>Polya's urn problem</td><td>power law distribution</td><td>$f_i \propto i^{(-1+1/(1-p))}$</td></tr>
<tr><td>$0 < \gamma < 1$</td><td rowspan="3">all bins grow at the same rate asymptotically</td><td rowspan="3">exponentially decreasing assuming $(*)$</td><td>$f_i \propto i^{-\gamma} e^{-Ki^{1-\gamma}/(1-\gamma)}$</td></tr>
<tr><td>$\gamma = 0$</td><td>$f_i \propto (K+1)^{-i}$</td></tr>
<tr><td>$\gamma < 0$</td><td>$f_i = O\left(((i-1)!)^{\gamma}/K^i\right)$</td></tr>
</table>

TABLE 1. **The distribution of bin sizes.**
f_i is the limit of the fraction of bins with i balls and $K = \frac{p}{1-p} \sum_{i=1}^{\infty} f_i \, i^{\gamma}$.

We next use Theorem 4.1 to approximate the asymptotic behavior of f_i when $\gamma < 1$, though we cannot evaluate K explicitly. If $0 < \gamma < 1$, then for large i,

$$\begin{aligned} f_i &\propto \prod_{j=2}^{i} \frac{(j-1)^{\gamma}}{K + j^{\gamma}} \propto i^{-\gamma} \prod_{j=1}^{i} \frac{j^{\gamma}}{K + j^{\gamma}} \\ &= i^{-\gamma} \prod_{j=1}^{i} \frac{1}{1 + \frac{K}{j^{\gamma}}} \propto i^{-\gamma} e^{-\sum_{j=1}^{i} K/j^{\gamma}} \\ &\propto i^{-\gamma} e^{-Ki^{1-\gamma}/(1-\gamma)}. \end{aligned}$$

If $\gamma = 0$, then $f_i \propto (K+1)^{-i}$. Finally, if $\gamma < 0$, then

$$f_i \propto \prod_{j=2}^{i} \frac{(j-1)^{\gamma}}{K + j^{\gamma}} = O\left(\prod_{j=2}^{i} \frac{(j-1)^{\gamma}}{K} \right) = O\left(\frac{((i-1)\,!)^{\gamma}}{K^i} \right).$$

□

5. Summary, problems and remarks

In Table 1 we summarize all the cases of the finite/infinite Polya process with positive/negative feedback. Many problems remain unresolved, several of which we mention here:

Problem 1: Prove assumptions $(*)$

For the infinite Polya process with exponent $\gamma < 1$, it would be of interest to

prove the assumptions (∗); in particular, is it true that the limit of

$$\sum_{i=1}^{\infty} f_{i,t} i^{\gamma}$$

exists and is finite? Recall that $f_{i,t}$ denotes the fraction of bins with i balls at time t. We note that for the case of $\gamma = 1$, this sum is just the average bin size, and so the limit exists and is $\frac{1}{p}$.

Problem 2: The rate of convergence

In preceding sections, we discussed the limit and convergence of the $f_{i,t}$'s. A natural problem is to analyze how fast convergence occurs. In the case of positive feedback, empirical results indicate that one bin very quickly becomes dominant. It would be desirable to provide estimates for the rate of convergence. Also, how does the rate of convergence vary if we alter the initial distribution of balls in bins?

Problem 3: Web models

In [7], Drinea et al. give an excellent exposition on the motivation of the problems of balls and bins with feedback and their relations to dynamic web-graph models. Many proposed web-graph models can be viewed as balls-and-bins problems with linear feedback ($\gamma = 1$) in the sense that a new page with one out-edge links to an existing page with probability proportional to its indegree. In [6], variations of web-graph models are introduced so that a new page with one out-edge links to an existing page with probability proportional to its indegree to the power γ. Although such generalizations can not generate power law distributions (except for the case of $\gamma = 1$), is it possible some areas of the Web may be similar to this more general model? What are the structural properties of such web graph models?

Problem 4: Random walks with feedback

A random walk on a graph can be viewed as a collection of balls-and-bins problems, one at each vertex: bins represent edges incident to a given vertex u, and balls in a bin represent instances when the random walk leaves u via that edge. If for some u, all the transition probabilities $p_{u,v}$ are equal, then the resulting balls-and-bins problem is our finite Polya process with exponent 0. However, the transition probabilities $p_{u,v}$ can instead depend on the number of times the random walk has traversed each edge leaving u. For example, for a given exponent γ, define p_{uv} to be proportional to $g(u,v)^{\gamma}$ where $g(u,v)$ is one plus the number of times that the edge uv has been traversed. The finite Polya process with k bins corresponds to the case of a random walk on a graph with one vertex and k loops. The case $\gamma = 1$ is the so-called *reinforced random*

walk, and was first introduced by Diaconis [5]. There have been a number of results on paths and trees for reinforced random walks. The reader is referred to an excellent survey of Pemantle [10] and his result with Volkov [11] on the convergence to five points on the infinite path. Numerous questions can be asked about such random walks with feedback, although it seems that very little is known.

References

[1] R. Albert and A.-L. Barabási, Statistical mechanics of complex networks, *Review of Modern Physics* **74** (2002), 47-97.

[2] B. Arthur, *Increasing returns and path dependence in the economy*, The University of Michigan press, 1994.

[3] B. Arthur, Y. Ermoliev, and Y. Kaniovski, On generalized urn schemes of the Polya kind, English translation in *cybernetics* 19: 61-71, 1963.

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

[5] P. Diaconis, Recent progress on de Finetti's notions of exchangeability, *Bayesian statistics* **3** (Valencia, 1987), 111-125.

[6] E. Drinea, M. Enachescu and M. Mitzenmacher, Variations on random graph models for the web, *Harvard Technical Report* TR-06-01.

[7] E. Drinea, A. Frieze and M. Mitzenmacher, Balls and bins models with feedback, SODA 2002.

[8] N. Johnson and S. Kotz, *Urn Models and Their Applications: an approach to Modern Discrete Probability Theory*, Wiley, New York, 1977.

[9] P. L. Krapivsky and S. Redner, Organization of growing random networks, *physica A* **281** (2000), 69-77.

[10] R. Pemantle, Vertex-reinforced random walks, *Probab. Theory Related Fields* **92** (1992), 117-136.

[11] R. Pemantle and S. Volkov, Vertex-reinforced random walk on **Z** has finite range, *Ann. Probab.* **27** (1999), 1368-1388.

[12] C. Shapiro and H. R. Varian, *Information Rules*, Harvard Business school Press, Boston, MA 1999.

[13] J. Spencer and N. Wormald, Explosive processes, manuscript.

Fan Chung, Department of Mathematics, UCSD, San Diego, CA 92093,
E-mail address: fan@ucsd.edu

Shirin Handjani, Center for Communications Research, IDA, 4320 Westerra Court, San Diego,CA 92121,
E-mail address: shandjan@math.ucsd.edu

Doug Jungreis, Center for Communications Research, IDA, 4320 Westerra Court, San Diego,CA 92121,
E-mail address: jungreis@ccrwest.org