A Scaling Result for Explosive Processes

M. Mitzenmacher*

Division of Engineering and Applied Sciences
Harvard University, Cambridge, MA 02138
michaelm@eecs.harvard.edu

R. Oliveira†, J. Spencer

Courant Institute of Mathematical Sciences
New York University, New York, NY 10012
{oliveira,spencer}@cims.nyu.edu

Submitted: Apr 7, 2003; Accepted: Feb 25, 2004; Published: Apr 13, 2004.
MR Subject Classifications: 60J20, 68R05

Abstract

We consider the asymptotic behavior of the following model: balls are sequentially thrown into bins so that the probability that a bin with n balls obtains the next ball is proportional to $f(n)$ for some function f. A commonly studied case where there are two bins and $f(n) = n^p$ for $p > 1$. In this case, one of the two bins eventually obtains a monopoly, in the sense that it obtains all balls thrown past some point. This model is motivated by the phenomenon of positive feedback, where the "rich get richer." We derive a simple asymptotic expression for the probability that bin 1 obtains a monopoly when bin 1 starts with x balls and bin 2 starts with y balls for the case $f(n) = n^p$. We then demonstrate the effectiveness of this approximation with some examples and demonstrate how it generalizes to a wide class of functions f.

1 Introduction

We consider the following balls and bins model: balls are sequentially thrown into bins so that the probability that a bin with n balls obtains the next ball is proportional to $f(n)$ for some function f. For example, a common case to study is when $f(n) = n^p$ for some constant $p > 1$. Specifically, we consider the case of two bins, in which case the state

*Supported in part by an Alfred P. Sloan Research Fellowship and NSF grants CCR-9983832, CCR-0118701, and CCR-0121154.

†Supported by a CNPq doctoral fellowship.

(x, y) denotes that bin 1 has x balls and bin 2 has y balls. In this case, the probability that the next ball lands in bin 1 is $\frac{x^p}{x^p+y^p}$.

This model is motivated by the phenomenon of *positive feedback*. In economics, positive feedback refers to a situation where a small number of companies compete in a market until one obtains a non-negligible advantage in the market share, at which point its share rapidly grows to a monopoly or near-monopoly. One loose explanation for this principle, commonly referred to as Metcalfe's Law, is that the inherent potential value of a system grows super-linearly in the number of existing users. Positive feedback also occurs in chemical and biological processes. For example, the above model is used in [4] to develop a model for neuron growth. For further examples, see [1]. Here we consider positive feedback between two competitors, with the strength of the feedback modeled by the parameter p, although our methods can also easily be applied to similar problems with more competitors.

It is known that for the model above that when $p > 1$ eventually one bin obtains a monopoly in the following sense: with probability 1 there exists a time after which all subsequent balls fall into just one of the bins [2, 7]. Given this limiting behavior, we now ask what is the probability that bin 1 will eventually obtain the monopoly starting from state (x, y). We provide an asymptotic analysis, based on examining the appropriate scaling of the system. This approach is reminiscent of techniques used to study phase transitions in random graphs, as well as other similar phenomena.

Our main result for the case where $f(n) = n^p$ and $p > 1$ can be stated as follows. Let $a = (x + y)/2$. We show that in the limit as a grows large, when $x = a + \frac{\lambda}{\sqrt{4p-2}}\sqrt{a}$, the probability that x obtains the monopoly converges to $\Phi(\lambda)$, where Φ is the cumulative distribution function for the normal distribution with mean 0 and variance 1. Throughout the paper, we treat quantities such as x as integers, as adding a ceiling or a floor does not change the asymptotic results.

The rest of the paper proceeds as follows. We first prove the theorem above for the specific case of $f(n) = n^p$ and $p > 1$. We show that the asymptotic approximation is extremely accurate with a pair of numerical examples. We follow with a more general statement that can be applied to a larger family of functions f. Related results and possible extensions are discussed in final section.

2 The case of $f(n) = n^p$

This section is devoted to the following theorem:

Theorem 1 *For the balls-and-bins process described above with $f(n) = n^p$ and $p > 1$, from the state (x, y) with $a = x+y$ and $x = a+\frac{\lambda}{\sqrt{4p-2}}\sqrt{a}$, the probability that bin 1 obtains the eventual monopoly is $\Phi(\lambda) + O(1/\sqrt{a})$.*

Proof: The argument utilizes an interesting embedding of the throwing process into time, apparently originally due to Rubin (as reported by Davis in [2]) and rediscovered by Spencer and Wormald [7]. With this embedding, if bin 1 has z balls at time t, it receives its next ball at a time $t + T_z$, where T_z is a random variable exponentially distributed with mean z^{-p}. Similarly, if bin 2 has z balls at time t, it receives its next ball at a time $t + U_z$, where U_z is a random variable exponentially distributed with mean z^{-p}. From the properties of the exponential distribution, we can deduce that this maintains the property that in any state (x, y), the probability that the next ball lands in bin 1 is proportional to x^p. Specifically, the probability that the minimum of the two exponentially distributed random variables T_x with mean x^{-p} and U_y with mean y^{-p} is T_x with probability $\frac{x^p}{x^p+y^p}$. Moreover, from the memorylessness of the exponential distribution, when a ball arrives at state (x, y) to bin 1 (respectively, bin 2), the time U_y (T_x) until the next ball arrives at bin 2 (bin 1) is still exponentially distributed with the same mean.

The *explosion time* for a bin is the time under this framework when a bin receives an infinite number of balls. If we begin at state (x, y) at time 0, the explosion time F_1 for bin 1 satisfies

$$F_1 = \sum_{j=x}^{+\infty} T_j = \sum_{j=a+\lambda\sqrt{a/(4p-2)}}^{+\infty} T_j.$$

Similarly, the explosion time F_2 for bin 2 is

$$F_2 = \sum_{k=y}^{+\infty} U_j = \sum_{k=a-\lambda\sqrt{a/(4p-2)}}^{+\infty} U_k.$$

Note that $E[F_1]$ and $E[F_2]$ are finite; indeed, the explosion time for each bin is finite with probability 1. Also, F_1 and F_2 are distinct with probability 1. This is easily seen by noting that $F_1 = F_2$ if and only if

$$T_x = \sum_{k=y}^{+\infty} U_k - \sum_{j=x+1}^{+\infty} T_j,$$

a probability 0 event. It is therefore evident that the bin with the smaller explosion time at some point obtains all balls thrown past some point, as first noted by Rubin in [2].

We first demonstrate that for sufficiently large a, F_1 and F_2 are approximately normally distributed. This would follow immediately from the Central Limit Theorem if the sum of the variances of the random variables T_j grew to infinity. Unfortunately,

$$\sum_{j=x}^{+\infty} \mathrm{Var}[T_j] = \sum_{j=x}^{+\infty} j^{-2p} < +\infty,$$

and hence standard forms of the Central Limit Theorem do not apply.

Fortunately, we may apply Esséen's inequality, a variation of the Central Limit Theorem, which can be found in, for example, [5][Theorem 5.4].

Lemma 1 *[Esséen's inequality] Let* $X_1, X_2, \ldots, X_n$ *be independent random variables with* $E[X_j] = 0$, $Var[X_j] = \sigma_j^2$, *and* $E[|X_j|^3] < +\infty$ *for* $j = 1, \ldots, n$. *Let* $B_n = \sum_{i=0}^n \sigma_j^2$, $F(x) = \mathbf{Pr}(B_n^{-1/2} \sum_{j=1}^n X_j < x)$, *and* $L = B_n^{-3/2} \sum_{j=1}^n E[|X_j|^3]$. *Then*

$$\sup_x |F(x) - \Phi(x)| \leq cL$$

for some universal constant c.

In our setting, let $X_j = T_{x+j-1} - (x+j-1)^{-p}$. We note that there are no problems applying Esséen's theorem to the infinite summations of our problem. Consider

$$F^x(z) = \mathbf{Pr}\left(\frac{\sum_{j=x}^{+\infty}(T_j - j^{-p})}{\sqrt{\sum_{j=x}^{+\infty} j^{-2p}}} < z \right).$$

That is, $F^x(z)$ is the probability that F_1, appropriately normalized to match a standard normal of mean 0 and variance 1, is less than or equal to z. Then we have

$$\sup_z |F^x(z) - \Phi(z)| \leq O(1/\sqrt{x}).$$

Hence $F^x(z)$ approaches a normal distribution as x grows large.

We also have

$$E[F_1] = \sum_{j=x}^{+\infty} E[T_j] = \sum_{j=x}^{+\infty} \frac{1}{j^p} = \frac{x^{1-p}}{p-1} + O(x^{-p}),$$

and

$$\mathrm{Var}[F_1] = \sum_{j=x}^{+\infty} \mathrm{Var}[T_j] = \sum_{j=x}^{+\infty} \frac{1}{j^{2p}} = \frac{x^{1-2p}}{2p-1} + O(x^{-2p}).$$

We wish to determine the probability that $F_1 - F_2 < 0$. Now $F_1 - F_2$ is (approximately) normally distributed with mean μ where

$$\mu = E[F_1] - E[F_2] = -2\frac{\lambda}{\sqrt{4p-2}} a^{1/2-p} + O(a^{-p})$$

and variance σ^2 where

$$\sigma^2 = \mathrm{Var}[F_1] + \mathrm{Var}[F_2] = \frac{2}{2p-1} a^{1-2p} + O(a^{-2p}).$$

Hence the probability that $F_1 - F_2 < 0$ is $\Phi(\lambda + O(1/\sqrt{a})) + O(1/\sqrt{a})$, which is just $\Phi(\lambda) + O(1/\sqrt{a})$. $\square$

3 Numerical Examples

We provide an example demonstrating the accuracy of Theorem 1 in Table 1. We consider initial states with 200 balls in the system, with the first bin containing between 101 and 110 balls. We estimate the exact probability that the first bin achieves monopoly as follows. We first calculate the exact distribution when there are 160,000 balls in the system for the case $p = 2$, using the recursive equations described in [3]. With this data, we make the very accurate approximation bin 1 eventually achieves monopoly if it has 53% of the balls at this point. We also apply symmetry for the remaining cases; if at this point bin 1 has $80{,}000 \leq k < 84{,}800$ balls with probability p_1 and bin 2 has k balls with probability $p_2 < p_1$, then bin 1 reaches monopoly at least 1/2 out of this $p_1 + p_2$ fraction of the time. This approach is sufficient to accurately determine the probability that the first bin eventually reaches monopoly to four decimal places. Comparing these results demonstrates the accuracy of the normal estimate. This accuracy is somewhat surprising, as our bound for the error of the estimate is $O(1/\sqrt{a})$; we suspect tighter provable bounds may be possible. Table 2 shows similar results for the case of $p = 1.5$. Here we calculate exactly the distribution with 640,000 balls in the system, use a 52% cutoff to estimate the probability of monopoly, and again use symmetry; the resulting numbers are correct to four decimal places. Again, the normal estimate provides a great deal of accuracy.

x	101	102	103	104	105
Calc.	0.5955	0.6870	0.7682	0.8361	0.8896
$\Phi(\lambda)$	0.5970	0.6883	0.7693	0.8370	0.8902
x	106	107	108	109	110
Calc.	0.9292	0.9569	0.9751	0.9863	0.9929
$\Phi(\lambda)$	0.9297	0.9572	0.9753	0.9865	0.9930

Table 1: A calculation vs. the asymptotic estimate of our theorem when $a = 100$ and $p = 2$.

x	101	102	103	104	105
Calc.	0.5794	0.6557	0.7261	0.7886	0.8419
$\Phi(\lambda)$	0.5793	0.6554	0.7257	0.7881	0.8413
x	106	107	108	109	110
Calc.	0.8854	0.9197	0.9456	0.9644	0.9775
$\Phi(\lambda)$	0.8849	0.9192	0.9452	0.9641	0.9772

Table 2: A calculation vs. the asymptotic estimate of our theorem when $a = 100$ and $p = 1.5$.

Feedback ($f = f(n)$)	Scale ($q = q(a)$)
$n^p \ln^\alpha n$	$\sqrt{\frac{a}{4p-2}}$
$n^p \ln n \ln\ln^\alpha n$	$\sqrt{\frac{a}{4p-2}}$
$n^{p+\ln^\alpha n}$	$\sqrt{\frac{a}{4(\alpha+1)\ln^\alpha a}}$

Table 3: Different feedback functions f and the asymptotic form of their corresponding scale functions q. Here p and α can be any constants for which the corresponding feedback function satisfies condition (1). The verification of the hypotheses of Theorem 2 is left to the reader.

4 A more general argument

We now prove a generalization of Theorem 1 to processes where the strength of feedback is modeled by a positive non-decreasing function $f : \mathbb{N} \to (0, +\infty)$. More precisely, the probability of bin 1 receiving the next ball when the current state of the system is (x, y) is $\frac{f(x)}{f(x)+f(y)}$. In this case we say that f is the *feedback function* of the process. It is known that any such f that satisfies

$$\sum_{n=1}^{+\infty} \frac{1}{f(n)} < +\infty \tag{1}$$

gives rise to a process for which with probability 1 one of the bins will receive all balls beyond a certain finite time [2, 7]. The aim of this Section is to characterize the asymptotic behavior of the probability of bin 1 achieving monopoly in a way that is analogous to Theorem 1.

Our main result is more easily expressed when f is defined over all the positive real numbers and is continuously differentiable, in which case we say that $q = q(a)$ is a *scale function* if $q(a) \sim \sqrt{\frac{a}{4a(\ln f)'(a)-2}}$ as $a \to +\infty$.[1] Theorem 2 states that if the process starts from initial state (x, y) with $a = \frac{x+y}{2}$, $x = a + \lambda q(a)$, and a large, the probability of monopoly by bin 1 is approximately $\Phi(\lambda)$. This is true whenever f satisfies certain technical conditions on its logarithmic growth rate. This result subsumes the $f(n) = n^p$ case treated in Theorem 1 (except for the error bounds), and although it is not completely general, it characterizes the scaling behavior of the monopoly probability in most interesting examples with sub-exponential growth, such as the ones given in Table 3 above.

The remainder of this Section is devoted to the proof of Theorem 2. We begin with a probabilistic result (Lemma 2) that provides sufficient conditions under which scaling behavior can be verified. The subsequent proof of Theorem 2 is analytic and consists of showing that the conditions of Lemma 2 are satisfied whenever some easily verifiable conditions on f hold.

[1]We shall sometimes speak of *the* scale function where in fact we are only referring to one of the many possible scale functions, all of which are asymptotically equivalent.

4.1 Sufficient conditions for scaling behavior

We generalize Theorem 1 with the following lemma.

Lemma 2 *Let* $\mathsf{mon}(x,y)$ *be the probability that bin* 1 *achieves monopoly (i.e. receives all balls beyond a certain time) in a balls-and-bins process started from state* (x,y) *whose feedback function* $f:\mathbb{N}\to(0,+\infty)$ *satisfies condition* (1). *Let*

$$S_r(n) = \sum_{j\geq n}\frac{1}{f(j)^r} \quad (n\in\mathbb{N}, r\in\{1,2,3\});$$
$$q_0(n) = f(n)\sqrt{\frac{S_2(n)}{2}} \ (n\in\mathbb{N}).$$

Choose some function $q=q(n)$ *and a fixed* $\lambda>0$. *Assume that there is a function* $0\leq \mathbf{er}(n)\ll 1$ *as* $n\to+\infty$ *such that*

$$0\leq\left|\frac{q(n)}{q_0(n)}-1\right|\leq\mathbf{er}(n); \tag{2}$$

$$0\leq\left|\frac{f(n\pm\lambda q(n))}{f(n)}-1\right|\leq\mathbf{er}(n); \tag{3}$$

$$0\leq\frac{S_3(n)}{S_2(n)^{3/2}}\leq\mathbf{er}(n). \tag{4}$$

Then

$$\mathsf{mon}(a+\lambda q(a), a-\lambda q(a)) = \Phi(\lambda) + O\left(\mathbf{er}(n)\right) \text{ as } a\to+\infty.$$

Proof: We essentially retrace the steps of the proof of Theorem 1. The exponential embedding technique again applies. We now assume that if bin 1 has z balls at time t receives its next ball at time $t+T_z$, where T_z is exponential with mean $f(z)^{-1}$, and we have similar random variables U_z for bin 2. As before, if we start from state (x,y), the elementary properties of the exponential distribution imply that the probability of the first arrival happening at bin 1 is

$$\mathbf{Pr}(T_x = \min\{T_x, U_y\}) = \frac{f(x)}{f(x)+f(y)}.$$

The memorylessness of the exponential implies that this same property holds for all subsequent arrivals, which are therefore distributed as the original balls-and-bins process. The *explosion times* F_1 and F_2 are again defined to be the times at which respectively bin 1 and bin 2 receive infinitely many balls in this modified framework. Hence

$$F_1 = \sum_{j=x}^{+\infty} T_j,$$

and F_1 is almost surely finite by condition (1):

$$E[F_1] = \sum_{j=x}^{+\infty} \frac{1}{f(j)} < +\infty.$$

Of course similar equations hold for F_2. It is clear that with probability 1 $F_1 \neq F_2$ and that bin 1 receives all balls beyond a certain time if and only if $F_1 < F_2$. Hence

$$\mathsf{mon}(x,y) = \mathbf{Pr}(F_1 < F_2). \tag{5}$$

We compute the asymptotics of $\mathsf{mon}(x,y)$ with $x = a + \lambda q(a)$ and $y = a - \lambda q(a)$ as $a \to +\infty$, where $\lambda > 0$ is fixed, under assumptions (2), (3) and (4). As in the previous proof, we use Esséen's Inequality (Lemma 1) to prove that F_1 and F_2 can both be approximated in distribution by Gaussian random variables with appropriate mean and variance. For F_1 this can be done by setting (using the notation of Lemma 1)

$$X_j = T_j - \frac{1}{f(x-1+j)} \; (j = 1, 2, 3, \dots)$$

and again noting that there are no problems in applying the Lemma to this infinite sequence of random variables. Since

$$\begin{aligned} \sum_{j=x}^{+\infty} \mathrm{Var}[X_j] &= \sum_{n=x}^{+\infty} \frac{1}{f(n)^2} &= S_2(x), \\ \sum_{j=x}^{+\infty} E[|X_j|^3] &= O\left(\sum_{n=x}^{+\infty} \frac{1}{f(n)^3}\right) &= O\left(S_3(x)\right) \end{aligned}$$

and by assumption (3), for $r = 2, 3$,

$$S_r(x) = S_r(a + \lambda q(a)) = (1 + O\left(\mathbf{er}(a)\right))S_r(a),$$

the error term in Esséen's inequality is of the order of

$$L = \frac{S_3(x)}{S_2(x)^{3/2}} = (1 + O\left(\mathbf{er}(a)\right))\frac{S_3(a)}{S_2(a)^{3/2}} = O\left(\mathbf{er}(a)\right).$$

This implies that the distribution of F_1 is $O\left(\mathbf{er}(a)\right)$-close to the distribution of a normal random variable with mean and variance given by

$$E[F_1] = S_1(x) \text{ and } \mathrm{Var}[F_1] = S_2(x) = (1 + O\left(\mathbf{er}(a)\right))S_2(a). \tag{6}$$

A analogous statement holds for F_2. As a result, the distribution of $F_1 - F_2$ is $O\left(\mathbf{er}(a)\right)$ close to that of a normal random variable with mean and variance given by

$$\mu = E[F_1] - E[F_2] = -\sum_{n=a-\lambda q(a)}^{a+\lambda q(a)-1} \frac{1}{f(n)} = -(1 + O\left(\mathbf{er}(a)\right))\frac{2\lambda q(a)}{f(a)},$$

$$\sigma^2 = \mathrm{Var}[F_1] + \mathrm{Var}[F_2] = (1 + O(\mathbf{er}(a)))2S_2(a).$$

It follows that

$$\mathsf{mon}(x, y) = \mathbf{Pr}(F_1 - F_2 < 0) = \Phi\left(-\frac{\mu}{\sigma}\right) + O(\mathbf{er}(a)).$$

By (2) and the definition of q_0

$$-\frac{\mu}{\sigma} = (1 + O(\mathbf{er}(a)))\frac{2\lambda q_0(a)}{f(a)\sqrt{2S_2(a)}} = (1 + O(\mathbf{er}(a)))\lambda.$$

The above finally implies

$$\mathsf{mon}(x, y) = \Phi((1 + O(\mathbf{er}(a)))\lambda) + O(\mathbf{er}(a)) = \Phi(\lambda) + O(\mathbf{er}(a)),$$

finishing the proof. □

4.2 The general result

Let $f : \mathbb{N} \to (0, +\infty)$ be a a feedback function (i.e. positive and non-decreasing). Letting $g(n) = \ln f(n)$, g can be easily extended to a piecewise affine function over all positive real numbers by linear interpolation. As a result, all feedback functions f can be extended to piecewise smooth functions on the positive real numbers. That is the class of functions to which Theorem 2 applies.

Theorem 2 *Assume that a function f is a positive, non-decreasing*[2]*, piecewise smooth function defined on the positive real numbers, and assume that it satisfies (1). Define $g(x) = \ln f(x)$ and $h(x) = xg'(x)$, where g' is the right derivative of g. Assume that*

$$\liminf_{x\to+\infty} h(x) > \frac{1}{2}, \quad \lim_{x\to+\infty} g'(x) = \lim_{x\to+\infty} \frac{h(x)}{x} = 0, \tag{7}$$

and also that there is a constant $C > 0$ such that for all $0 < \epsilon < 1/2$ and all x big enough

$$\sup_{x\le t\le x^{1+\epsilon}} \left|\frac{h(t)}{h(x)} - 1\right| \le C\epsilon. \tag{8}$$

It then holds that $\sqrt{\frac{a}{4h(a)-2}}$ is the scale function of the balls-and-bins process with feedback function f. That is, if

$$q(a) \sim \sqrt{\frac{a}{4h(a) - 2}} \text{ as } a \to +\infty,$$

then for any fixed $\lambda > 0$ the probability of monopoly by bin 1 in such a process started from state $(x, y) = (a + \lambda q(a), a - \lambda q(a))$ converges to $\Phi(\lambda)$ as $a \to +\infty$.

[2] Condition (7) implies that $f = f(x)$ is in fact increasing in x for x big enough.

Proof: We shall check that the conditions of Lemma 2 are satisfied. The crucial step in checking these conditions is to estimate $S_2(n)$ and $S_3(n)$, which we accomplish by evaluating corresponding integrals. Let $r \geq 2$ and define

$$I_r(a) = \int_a^{+\infty} \frac{dx}{f(x)^r} = \int_a^{+\infty} \frac{dx}{e^{rg(x)}}.$$

In what follows we will prove that

$$S_r(a) \sim I_r(a) \sim \frac{a}{(rh(a)-1)f(a)^r} \text{ as } a \to +\infty.$$

By integration by parts,

$$I_r(a) = \frac{x}{e^{rg(x)}}\Big]_{x=a}^{x=+\infty} + r\int_a^{+\infty} \frac{xg'(x)\,dx}{e^{rg(x)}} = -\frac{a}{f(a)^r} + r\int_a^{+\infty} \frac{h(x)\,dx}{e^{rg(x)}}.$$

Here we have used the fact that

$$f(x)^r \gg x \text{ as } x \to +\infty \text{ for } r \geq 2, \tag{9}$$

which can be deduced from the fact that $\liminf_{x\to+\infty} h(x) > \frac{1}{2}$. We now make use of the following claim, which we prove subsequently.

Claim 1 *As* $a \to +\infty$

$$\int_a^{+\infty} \frac{h(x)\,dx}{e^{rg(x)}} \sim h(a)\int_a^{+\infty} \frac{dx}{e^{rg(x)}} = h(a)I_r(a). \tag{10}$$

□

Claim 1 implies that $a \to +\infty$

$$I_r(a) = -\frac{a}{f(a)^r} + (1+o(1))rh(a)\int_a^{+\infty} \frac{dx}{e^{rg(x)}} = -\frac{a}{f(a)} + (1+o(1))rh(a)I_r(a).$$

Assumption (7) tells us that $rh(a) > 1$ for $r \geq 2$ and a big enough. This permits us to write

$$I_r(a) = (1+o(1))\frac{a}{(rh(a)-1)f(a)^r}.$$

Since by (7), $a \gg h(a)$, we have

$$I_r(a) \gg \frac{1}{f(a)^r}.$$

Noting that $|S_r(a) - I_r(a)| \leq \frac{1}{f(a)^r}$, we can finally conclude

$$S_r(a) \sim I_r(a) \sim \frac{a}{(rh(a)-1)f(a)^r} \text{ as } a \to +\infty \ (r \geq 2). \tag{11}$$

This gives us the asymptotic form of S_2 and S_3 as in Lemma 2. Moreover, we can compute

$$q_0(n) = f(n)\sqrt{\frac{S_2(n)}{2}} \sim \sqrt{\frac{n}{4h(n)-2}}.$$

All that remains to be shown is that the assumptions of Lemma 2 hold in this case. For convenience we simply show that $\mathbf{er}(a) = o(1)$. To this end, we let

$$q(n) \sim \sqrt{\frac{n}{4h(n)-2}} \text{ as } n \to +\infty,$$

and note that this guarantees the validity of (2). To finish the proof, we show that as $a \to +\infty$,

$$S_3(a) \ll S_2(a)^{3/2}; \tag{12}$$

$$\forall \lambda > 0 \; f(a \pm \lambda q(a)) \sim f(a). \tag{13}$$

The first of these equations follows from (11) and equation (7) ($\frac{a}{h(a)} \gg 1$).

$$S_3(a) \sim \frac{a}{(3h(a)-1)f(a)^3} \ll S_2(a)^{3/2} \sim \frac{1}{f(a)^3}\left(\frac{a}{2h(a)-1}\right)^{3/2}.$$

To prove (13), fix an arbitrary $\lambda > 0$. By the definition of h,

$$|g(a \pm \lambda q(a)) - g(a)| \leq \left| \int_a^{a \pm \lambda q(a)} h(t)\frac{dt}{t} \right| \leq \ln\left(\frac{a+\lambda q(a)}{a-\lambda q(a)}\right)\left\{\sup_{a-\lambda q(a) \leq t \leq a+\lambda q(a)} h(t)\right\}.$$

Since $q(a) = O\left(\sqrt{a}\right)$, (8) implies

$$\sup_{a-\lambda q(a) \leq t \leq a+\lambda q(a)} h(t) \sim h(a).$$

We conclude (again using $q(a) = O\left(\sqrt{a}\right)$) that

$$|g(a \pm \lambda q(a)) - g(a)| \sim h(a)\ln\left(\frac{a+\lambda q(a)}{a-\lambda q(a)}\right) = O\left(\frac{h(a)}{a}q(a)\right) = O\left(\sqrt{\frac{h(a)}{a}}\right) = o(1),$$

because $a \gg h(a)$ by (7). Hence

$$\frac{f(a \pm \lambda q(a))}{f(a)} = e^{g(a \pm \lambda q(a)) - g(a)} = e^{o(1)}.$$

This proves (13) and finishes the proof. $\square$

To conclude, we now prove Claim 1.

Proof: [of Claim 1] We first show that for any fixed $0 < \epsilon < \frac{1}{2}$, as $a \to +\infty$,

$$\frac{\int_a^{a^{1+\epsilon}} \frac{h(x)\,dx}{e^{rg(x)}}}{\int_a^{+\infty} \frac{h(x)\,dx}{e^{rg(x)}}} \sim 1. \tag{14}$$

A change of variables permits us to rewrite

$$\int_{a^{1+\epsilon}}^{+\infty} \frac{h(x)\,dx}{e^{rg(x)}} = (1+\epsilon) \int_a^{+\infty} \frac{h(u^{1+\epsilon})u^{\epsilon}\,du}{e^{rg(u^{1+\epsilon})}}. \tag{15}$$

Equation (8) implies that for all u big enough, $h(u^{1+\epsilon}) \leq (1+C\epsilon)h(u)$. Moreover, (7) allows us to choose an a such that $h(u) \geq h_0 > \frac{1}{2}$ for all $u \geq a$, which implies

$$g(u^{1+\epsilon}) - g(u) = \int_u^{u^{1+\epsilon}} g'(u)du \geq \inf_{t \geq a} h(t) \int_u^{u^{1+\epsilon}} \frac{du}{u} = h_0 \epsilon \ln u.$$

We therefore find

$$e^{rg(u^{1+\epsilon})} \geq u^{rh_0\epsilon} e^{rg(u)}. \tag{16}$$

Also note $rh_0\epsilon > \epsilon$.

Plugging this into (15) yields the following estimate as $a \to +\infty$:

$$\int_{a^{1+\epsilon}}^{+\infty} \frac{h(x)\,dx}{e^{rg(x)}} \leq (1+\epsilon)(1+C\epsilon) \int_a^{+\infty} \frac{h(u)u^{\epsilon}\,du}{e^{rg(u)}u^{rh_o\epsilon}} = O\left(a^{\epsilon - rh_0\epsilon}\right) \int_a^{+\infty} \frac{h(u)\,du}{e^{rg(u)}}.$$

By (16), this implies

$$\int_a^{+\infty} \frac{h(x)\,dx}{e^{rg(x)}} \sim \int_a^{a^{1+\epsilon}} \frac{h(x)\,dx}{e^{rg(x)}}$$

as stated. Now note that, by assumption (8) on h,

$$(1-C\epsilon)h(a)\int_a^{a^{1+\epsilon}} \frac{dx}{e^{rg(x)}} \leq \int_a^{a^{1+\epsilon}} \frac{h(x)\,dx}{e^{rg(x)}} \leq (1+C\epsilon)h(a)\int_a^{a^{1+\epsilon}} \frac{dx}{e^{rg(x)}}$$

and by a similar reasoning as above

$$\int_a^{a^{1+\epsilon}} \frac{dx}{e^{rg(x)}} \sim \int_a^{+\infty} \frac{dx}{e^{rg(x)}}.$$

Putting these facts together finishes the proof of the claim. $\square$

5 Final remarks

We have provided a full description of scaling behavior of the probability of monopoly for a broad class of feedback functions satisfying condition (1), which corresponds to $p > 1$ in the $f(n) = n^p$ case. One is tempted to ask whether similar results hold in the $0 < p \leq 1$ range; in particular, it seems especially intriguing that the scale function

$$q(a) = \sqrt{\frac{a}{4p-2}}$$

for the $p > 1$ case can in fact be defined for all $p > 1/2$. It turns out [4] that any feedback function f satisfying

$$\sum_{n=1}^{+\infty} \frac{1}{f(n)^2} < +\infty \tag{17}$$

yields a process such that with probability 1, one of the bins has more balls than the other at all sufficiently large times. In forthcoming work, Oliveira and Spencer [6] prove that, if $f(n) = n^p$, $p > 1/2$, the probability a bin obtains *eventual leadership* has a standard Gaussian limit precisely at the $\lambda\sqrt{\frac{a}{4p-2}}$ scale, and similar results hold in the general context of Theorem 2 if assumption (1) is dropped. They also show that the limit of the *leadership* probability, which is defined to be the probability that bin 1 has more balls at all subsequent times, is $2\Phi(\lambda) - 1$ under the same scaling.

Many other natural questions remain open. For instance, are our methods applicable to related non-linear models for Web graphs [3]? It seems likely that this problem requires improvements on the error bounds for Gaussian approximation, and our numerical data suggests that this is indeed possible. However, it is also conceivable that large deviation bounds are enough for treating many related problems. Finally, direct combinatorial proofs (i.e. without resort to the exponential random variables) of the current results presented here would also be of great interest.

References

[1] B. Arthur. **Increasing Returns and Path Dependence in the Economy**. The University of Michigan Press, 1994.

[2] B. Davis. Reinforced Random Walks. *Probability Theory and Related Fields*, 84:203-229, 1990.

[3] E. Drinea, A. Frieze, and M. Mitzenmacher. Balls and Bins Models with Feedback. In *Proceedings of 13th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 308-315, 2002.

[4] K. Khanin and R. Khanin. A probabilistic model for establishment of neuron polarity. Technical Report HPL-BRIMS-2000-16, June 2000.

[5] V. Petrov. **Limit Theorems of Probability Theory**. Oxford University Press, 1995.

[6] R. Oliveira and J. Spencer. In preparation.

[7] J. Spencer and N. Wormald. Explosive processes. Draft manuscript.