The Hiring Problem and Lake Wobegon Strategies*

Andrei Z. Broder† Adam Kirsch‡ Ravi Kumar§ Michael Mitzenmacher¶
Eli Upfal‖ Sergei Vassilvitskii**

October 3, 2007

Abstract

We introduce the *hiring problem*, in which a growing company continuously interviews and decides whether to hire applicants. This problem is similar in spirit but quite different from the well-studied secretary problem. Like the secretary problem, it captures fundamental aspects of decision making under uncertainty and has many possible applications.

We analyze natural strategies of *hiring above the current average*, considering both the mean and the median averages; we call these *Lake Wobegon strategies*. Like the hiring problem itself, our strategies are intuitive, simple to describe, and amenable to mathematically and economically significant modifications. We demonstrate several intriguing behaviors of the two strategies. Specifically, we show dramatic differences between hiring above the mean and above the median. We also show that both strategies are intrinsically connected to the lognormal distribution, leading to only very weak concentration results, and the marked importance of the first few hires on the overall outcome.

1 Introduction

One of the most famous combinatorial mathematical questions is the secretary problem (also known as the marriage problem, the optimal stopping problem, the Sultan's dowry problem, and by several other names): applicants for a secretarial position are interviewed in a random order, and the relative rank of an applicant compared to previous applicants is revealed. After an interview, the applicant must either be accepted or rejected before the next interview. The goal is to maximize the

*A preliminary version of this paper appears in [2].

†Yahoo! Research, Sunnyvale, CA. Email: broder@yahoo-inc.com

‡Harvard School of Engineering and Applied Sciences, Cambridge, MA. Work done in part while visiting Yahoo! Research. Supported in part by an NSF Graduate Research Fellowship, NSF Grant CCR-0121154, and a grant from Yahoo! Research. Email: kirsch@eecs.harvard.edu

§Yahoo! Research, Sunnyvale, CA. Email: ravikumar@yahoo-inc.com

¶Harvard School of Engineering and Applied Sciences, Cambridge, MA. Work done in part while visiting Yahoo! Research. Supported in part by NSF Grant CCR-0121154 and a grant from Yahoo! Research. Email: michaelm@eecs.harvard.edu

‖Department of Computer Science, Brown University, Providence, RI. Work done in part while visiting Yahoo! Research. Supported in part by NSF Award DMI-0600384, ONR DEPSCOR Award N000140610607, and a grant from Yahoo! Research. Email: eli@cs.brown.edu

**Yahoo! Research, New York, NY. Work done in part while in the Department of Computer Science at Stanford University. Supported in part by a Microsoft Research Live Labs fellowship and NSF Grant ITR-0331640. Email: sergei@yahoo-inc.com

probability of accepting the best applicant, and the problem is to design a strategy that maximizes this probability.

Since its introduction in the 1960's [6], the secretary problem has been the subject of dozens of papers. (See, e.g., [1, 4, 5, 7, 9–11, 16, 17, 19] and the references therein.) This simple abstraction highlights the problem of choosing the best of a set of sequentially presented random variables, and thereby captures fundamental issues and inevitable tradeoffs related to making irrevocable decisions under an uncertain future. As such, it spans multiple scientific disciplines, such as mathematics, economics, and computer science. Furthermore, the basic form of the problem is easily stated, understood, and amenable to many variations that capture particular settings.

We introduce a problem in the same spirit that also captures basic tradeoffs in the face of uncertainty. To honor mathematical history and the connection to the secretary problem, we call it the *hiring problem.* In our setting, a small, nimble, start-up company that intends to grow into a huge, evil, multinational corporation begins hiring employees. The company wants to ensure a high quality of staff by consistently improving average employee quality. On the other hand, the company needs working bodies, and it cannot simply wait for the best candidate to come along. As in the secretary problem, applicants are interviewed, and the decision is immediate. In contrast to the secretary problem, however, hiring is done continuously, with no fixed limit in mind. The basic tradeoff in this setting is between the rate at which employees are hired and their quality.

Like the secretary problem, the hiring problem captures a fundamental issue that arises in many applications where one must make decisions under uncertainty. We emphasize that the hiring problem is as much about a company hiring employees as the secretary problem is about a person hiring a secretary. (That is, only tangentially.) Rather, the general problem statement is meant to give insight into a general mathematical question with many possible applications.

While one could consider many strategies that balance the rate of hiring and the quality of the resulting hires, we analyze two natural strategies, that, following Peter Norvig [14], we denote as *Lake Wobegon strategies*[1]: hire applicants that are better than the average employee you already have, where by average we refer to either the mean or the median. Such strategies are not entirely theoretical: in [14] it is claimed that Google actually uses hiring above the mean, and a small simulation is presented to show that it leads to higher average quality than hiring above the minimum, even in the presence of noise. (Ignoring noise, this follows easily from our results: the average quality when hiring above the mean converges to 1, while when hiring above the minimum, it converges to $(1-\mu)/2$ where μ is the initial minimum quality.) Additionally, at least one author of this paper has heard of this strategy in the setting of tenure decisions within a department: to improve a department, only tenure junior faculty member whose quality lies above the current average tenured faculty. As we explain further below, the intuition behind this approach is that it leads to consistent improvement in employee quality.

One initial issue to be dealt with regards how applicant scores are determined. In this paper, we consider applicant scores to be interpreted as arbitrarily small quantiles of some predictive measure of an applicant's contribution to the company (for example, IQ, although that is unlikely to be the most desirable measure in practice). Thus, for reasons we explain more clearly below, we model scores as uniformly distributed on the interval $(0, 1)$. Our notion of an average employee is one

[1] As explained in the Wikipedia, Lake Wobegon is a fictional town where "the women are strong, the men are good looking, and all the children are above average." The Lake Wobegon effect in psychology refers to the tendency of people to overestimate their abilities, so that a large majority of people believe themselves to be above average. This term matches our strategies in which every employee, at least at the time they are hired, is above average.

whose quality score is either the mean or the median of all employees.

We find several interesting behaviors for these processes, using both mathematical analysis and simulation, including the following:

- Hiring above the median and hiring above the mean lead to greatly different behaviors.
- Both processes are intrinsically connected to lognormal distributions, leading to only very weak concentration bounds on the average quality.
- Both processes exhibit strong dependence on the initial conditions; in economic terms, this means the first few hires of your start-up can have a tremendous effect!

We emphasize that this paper represents just a first attempt to study this problem (and indeed some of our results are cut here for lack of space). Given our initial findings, we expect there to be further work on variations of and alternative strategies for the hiring problem in the future. We discuss some natural directions in the conclusion.

2 Introducing the Model

2.1 Definitions and Motivation

We suppose that we are interviewing applicants for positions at a company, one-by-one. Each applicant has a *quality score* Q_i, and we assume that these scores are independent with common distribution $Q \sim \text{Unif}(0,1)$, where this notation is read as Q is distributed according to the uniform distribution on $(0,1)$. While interviewing an applicant, we observe his quality score; for the strategies we study, we choose to hire an applicant if and only if his quality score is at least the *average score* of the current employees, for an appropriate notion of average. We assume henceforth that we start with one employee with some particular quality $q \in (0,1)$. We do this because if the first employee has quality score $U \sim \text{Unif}(0,1)$, then the number of interviews needed to hire a second employee with score at least U is geometrically distributed with mean $1/(1-U)$, and hence is $\mathbf{E}[1/(1-U)] = \infty$. We avoid this undesirable situation by conditioning on the first employee having fixed quality $0 < q < 1$.

The choice of $Q \sim \text{Unif}(0,1)$ has a natural interpretation. If Q instead has some non-uniform continuous distribution, then we can define an alternate notion of the quality of a candidate by $F^{-1}(Q_i)$, where F is the cumulative distribution function of Q. Then the $F^{-1}(Q_i)$'s are independent $\text{Unif}(0,1)$ random variables, and $F^{-1}(Q_i)$ is a natural measure of quality: it corresponds to the applicant's quality presented as an arbitrarily small quantile in the original metric. Since this transformation can be performed regardless of the original distribution of quality scores, the assumption that $Q_i \sim \text{Unif}(0,1)$ is not only very realistic, it is arguably the best one to use. Note that this transformation preserves ordering, so that hiring above the median leads to the same decision process regardless of the initial measure. It is perhaps less clear that the strategy of hiring above the mean is natural in the setting of such a transformation, although we still suspect it to be relevant to practical scenarios, if only because the mean is an intuitively simpler concept for most people than the median.

We emphasize that throughout our analysis of the Lake Wobegon strategies, we assume that the exact value of Q_i is determined during the interview process. Of course, it is natural to try to extend the model so that, during an interview, we obtain only an estimate $\hat{Q}_i$ of Q_i. We consider this variation of the model in Section 6.3.

2.2 Baseline Strategies

Before studying our Lake Wobegon strategies, it is worth considering two other seemingly natural strategies and pointing out their potential flaws. We emphasize, however, that all strategies involve tradeoffs between the quality and speed of hires, so there is no single best strategy.

Let us first consider what happens if we choose to only hire applicants with quality scores above a pre-specified threshold t. It should be clear that this strategy leads to a collection of employees with scores uniformly distributed between t and 1. With a large value of t, this strategy guarantees high quality from the beginning. However, because the threshold is fixed, this strategy does not lead to continual improvement as the size of the company grows. Perhaps more problematically, since the rate at which employees are hired is $1 - t$, fixing t requires us to make a very stark tradeoff between the overall quality of our employees and the rate of our company's growth (which is particularly important early on, when the company is small). This weakness in the threshold strategy seems difficult to overcome. In contrast, the Lake Wobegon strategies generally allow faster hiring when the company is small, with increasingly selective hiring as the company grows.

It is also natural to consider the strategy where we start with a single employee with some particular quality score q and only hire applicants whose quality scores are higher than the scores of all current employees. We refer to this as the *Max* strategy, and we sketch an analysis of it here because it introduces some important ideas that appear in our analyses of the Lake Wobegon strategies. For convenience, rather than considering the quality score Q_i of the ith hire, we consider the *gap* $G_i = 1 - Q_i$ between the score and 1, with $G_0 = g = 1 - q$ denoting the gap for the first employee. By conditioning, we have that G_i is uniformly distributed on $[0, G_{i-1}]$, so that $\mathbf{E}[G_i \mid G_{i-1}] = G_{i-1}/2$, and hence inductively we find $\mathbf{E}[G_n] = g/2^n$. Thus, the expected size of the gap shrinks exponentially as the number of hires grows.

On the other hand, we also have the multiplicative representation $G_i = G_{i-1}U_i$, where the U_i are independent uniform $(0, 1)$ random variables. This equation shows that the expected number of interviews required between any two hirings is actually infinite, for reasons similar to those given in Section 2.1. For example, since $G_0 = g$, we have $G_1 = gU_1$ and hence the expected number of interviews between the first and second hirings is $\mathbf{E}[1/G_1] = \mathbf{E}[1/U_1]/m = \infty$. While one could conceivably avoid this problem by changing the model in various ways, this issue certainly highlights a key problem with this strategy: large lags between hires.

Turning our attention away from the time between hires and back to the employees' quality scores, we can also use the multiplicative representation to determine the limiting distribution of G_n. This multiplicative process is best handled by taking the logarithm of both sides, from which we obtain $\ln G_i = \ln G_{i-1} + \ln U_i$, or inductively, $\ln G_n = \ln g + \sum_{i=1}^{n} \ln U_i$. The summation $\sum_{i=1}^{n} \ln U_i$ has an approximately normal distribution by the central limit theorem. More formally, $\mathbf{E}[\ln U_i] = -1$ and $\mathbf{Var}[\ln U_i] = 1$, and hence $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1 + \ln U_i)$ converges to $\mathrm{N}(0, 1)$, the normal distribution with mean 0 and variance 1, as $n \to \infty$. Thus, for large n, we have

$$\begin{aligned} \ln G_n &= \ln g + \sum_{i=1}^{n} \ln U_i \\ &\approx \ln g + \mathrm{N}(-n, n) = \mathrm{N}(\ln g - n, n), \end{aligned}$$

and so G_n has an (approximately) *lognormal* distribution (see, e.g., [12]). Interestingly, the above equation implies that the median value for G_n is approximately g/e^n for large n, since the normal distribution is symmetric around its mean. The mean and the median of G_n are therefore vastly

different (recall that $\mathbf{E}[G_n] = g/2^n$), and we can see that the distribution of G_n is highly skewed for large n. This phenomenon is a recurring theme throughout our analysis.

3 Hiring Above the Mean

We now move on to studying our Lake Wobegon strategies. We begin with hiring above the mean, for which the analysis is a bit simpler.

Let A_i denote the average quality after i hires, with $A_0 = q$ being the quality of the initial employee (so that A_i refers to the average of $i+1$ employees). The following basic convergence result is straightforward and given without proof, but we state it so that we are clear that all of the random variables that we use in our analysis are well-defined.

Proposition 3.1. *With probability 1, we hire infinitely many candidates and* $\lim_{i\to\infty} A_i = 1$.

3.1 Analysis of Expectations

To quantify the rate at which we hire applicants and the rate of convergence of the A_i's, we proceed by studying the *gap* sequence given by $G_i = 1 - A_i$, which converges to 0 almost surely. In what follows, we let $g = G_0 = 1 - q$ denote the initial gap. In this setting, we have a pleasant form for G_t for any $t \geq 0$, as given by the following lemma:

Lemma 3.1. *For any $t \geq 0$, the conditional distribution of G_{i+t} given G_i is the same as that of $G_i \prod_{j=1}^{t}[1 - U_j/(i+j+1)]$, where the U_j's are independent* $\mathrm{Unif}(0,1)$ *random variables.*

Proof. We proceed by induction on $t \geq 0$. For $t = 0$, the result is trivial. Now, suppose that $t > 0$ and that the claim holds for $t-1$. Then conditioned on $G_i, \ldots, G_{i+t-1}$, the quality score of the $(i+t)$th hired candidate is clearly $\mathrm{Unif}(1 - G_{i+t-1}, 1) \sim 1 - G_{i+t-1}U_t$. Thus, conditioned on $G_1, \ldots, G_{i+t-1}$, we have

$$\begin{aligned} G_{i+t} = 1 - A_{i+t} &\sim 1 - \frac{(i+t)A_{i+t-1} + (1 - G_{i+t-1}U_t)}{i+t+1} \\ &= 1 - \frac{(i+t)(1 - G_{i+t-1}) + (1 - G_{i+t-1}U_t)}{i+t+1} \\ &= G_{i+t-1}\left(1 - \frac{1-U_t}{i+t+1}\right) \sim G_{i+t-1}\left(1 - \frac{U_t}{i+t+1}\right), \end{aligned}$$

completing the proof. □

Using Lemma 3.1, we can derive formulas for the expected gap and the expected number of interviews after hiring n people. We then compute the asymptotics of these quantities directly from these formulas.

Proposition 3.2.

$$\mathbf{E}[G_n] = g\prod_{j=1}^{n}\left(1 - \frac{1}{2(j+1)}\right) = \Theta\left(1/\sqrt{n}\right).$$

Proof. By the Taylor series for $\ln(1+x)$ for $|x|<1$,

$$\begin{aligned}\mathbf{E}[G_n] &= g\prod_{j=1}^{n}\left(1-\frac{1}{2(j+1)}\right)\\ &= g\exp\left[\sum_{j=1}^{n}\ln\left(1-\frac{1}{2(j+1)}\right)\right]\\ &= g\exp\left[-\frac{1}{2}\ln n+\Theta(1)\right]=\Theta\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

□

Proposition 3.3. *Let T_n be the number of candidates that are interviewed before n are hired.*

$$\mathbf{E}\left[T_n\right]=\frac{1}{g}\sum_{i=1}^{n}\prod_{j=1}^{i-1}(j+1)\ln(1+1/j)=\Theta\left(n^{3/2}\right).$$

Proof. By Lemma 3.1,

$$\begin{aligned}\mathbf{E}\left[\frac{1}{G_n}\right] &= \frac{1}{g}\prod_{i=1}^{n}\mathbf{E}\left[\frac{1}{1-\mathrm{Unif}(0,1)/(i+1)}\right]\\ &= \frac{1}{g}\prod_{i=1}^{n}(i+1)\ln(1+1/i)\\ &= \frac{1}{g}\prod_{i=1}^{n}(i+1)\left(\frac{1}{i}-\frac{1}{2i^2}+\Theta\left(\frac{1}{i^3}\right)\right)\\ &= \frac{n+1}{g}\prod_{i=1}^{n}\left(1-\frac{1}{2i}+\Theta\left(\frac{1}{i^2}\right)\right)\\ &= \frac{n+1}{g}\exp\left[\sum_{i=1}^{n}\ln\left(1-\frac{1}{2i}+\Theta\left(\frac{1}{i^2}\right)\right)\right]\\ &= \frac{n+1}{g}\exp\left[\sum_{i=1}^{n}\left(-\frac{1}{2i}+\Theta\left(\frac{1}{i^2}\right)\right)\right]\\ &= \frac{n+1}{g}\exp\left[-\frac{1}{2}\ln n+\Theta(1)\right]\\ &= \Theta\left(\sqrt{n}\right).\end{aligned}$$

Let T_j' denote the number of candidates interviewed between the $(j-1)$st hire and the jth hire. Then the conditional distribution of T_j' given G_{j-1} is geometric with parameter G_{j-1}, and so $\mathbf{E}[T_j']=\mathbf{E}[1/G_{j-1}]$. Therefore,

$$\mathbf{E}[T_n]=\sum_{i=1}^{n}\mathbf{E}\left[\frac{1}{G_{i-1}}\right]=\sum_{i=1}^{n}\Theta\left(\sqrt{i}\right)=\Theta\left(n^{3/2}\right).$$

□

It is worth noting that the initial starting gap $g = G_0$ has a *multiplicative* effect on the expected gap and the expected number of interviews, as seen in Proposition 3.2 and Proposition 3.3. This would still be true even if we started with more than one employee; that is, generally, initial differences in the gap lead to multiplicative differences in these expectations, demonstrating the importance of the initial hires on the growth of an organization under this strategy.

Also, it is worth noting that after $n^{3/2}$ candidates, the expected value for the mean gap for the best n candidates is $\Theta(n^{-1/2})$. Hence hiring above the mean is, in this sense, within a constant factor of optimal.

3.2 Convergence and Concentration

We have now given results concerning the expectation of the number of interviews to hire n people and the expectation of the resulting gap after hiring n people. While these results are already useful, there is more that we can say. In this section, we show that the distribution of the gap weakly converges to a lognormal distribution. By weak convergence, we mean that the body of the distribution converges to a lognormal distribution (under suitable initial conditions), but there may be larger error at the tails. In fact, our simulation results demonstrate this, as we show in Section 5.

In order to give stronger bounds regarding the behavior of the gap at the tails of the distribution, we provide a martingale-based concentration argument. It is important to note that this argument does not give the extremely strong concentration around the mean that usually arises in applications of this technique; indeed, it clearly cannot, since the body of the distribution is converging to the heavy-tailed lognormal distribution. Still, we find substantially better concentration at the tails of the distribution through our martingale-based argument than can be obtained using Chebyshev's inequality or other weaker, standard approaches.

We begin our convergence arguments with a technical lemma.

Lemma 3.2. *Let $U \sim \text{Unif}(0,1)$ and $j \geq 1$. Then*

$$\mathbf{E}[\ln(1 - U/j)] = -1 - (j-1)\ln(1 - 1/j) = -1/2j + o(1/j)$$

$$\mathbf{Var}[\ln(1 - U/j)] = 1 - (j-1)j\ln^2(1 - 1/j) = 1/12j^2 + o(1/j^2)$$

$$\mathbf{E}\left[(\ln(1 - U/j) - \mathbf{E}[\ln(1 - U/j)])^4\right] = 1/80j^4 + o(1/j^4)$$

Proof. This result follows easily from integration and the Taylor series of $\ln(1+x)$ for $|x| < 1$. □

Proposition 3.4. *$\ln G_n - \mathbf{E}[\ln G_n]$ converges to some random variable G almost surely and in mean square as $n \to \infty$.*

Proof. In light of Lemma 3.1, we may abuse the definition of our probability space and write $G_n = g\prod_{i=1}^{n}(1 - U_i/(i+1))$, for independent $\text{Unif}(0,1)$ random variables U_1, U_2, ... Letting

$$Y_i = \ln(1 - U_i/(i+1)) - \mathbf{E}[\ln(1 - U_i/(i+1))]$$

for $i \geq 1$ gives $Z_n \triangleq \ln G_n - \mathbf{E}[\ln G_n] = \sum_{i=1}^{n} Y_i$. Since $\mathbf{E}[Y_i] = 0$ and the Y_i's are independent, the sequence $Z_1, Z_2, \ldots$ is a zero-mean martingale. Furthermore, $\mathbf{E}[Z_n^2] = \sum_{j=1}^{\infty}\mathbf{Var}[Y_j] = O(1)$ by Lemma 3.2. We may now apply a variant of the martingale convergence theorem (see, e.g. [8, Theorem 7.8.1]) to obtain the desired result. □

Given that $\ln G_n$ is the sum of independent random variables, one might expect that it can be asymptotically approximated by a normal distribution, implying that G_n can be asymptotically approximated by a lognormal distribution. Some care must be taken however, since $\ln G_n$ is not the sum of identically distributed independent random variables, and the means and variances of the summands shrink rather quickly. These facts require that we use some care in stating the result, and that we use a strong form of the central limit theorem commonly known as the Berry–Esséen inequality.

Lemma 3.3 ([15, Theorem 5.4])**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables with $\mathbf{E}[X_j] = 0$ and $\mathbf{E}[|X_j|^3] < \infty$ for $j = 1, \ldots, n$. Let $B_n = \sum_{j=1}^n \mathbf{Var}[X_j]$. There is a constant c such that*

$$\sup_x \left| \mathbf{Pr}\left(B_n^{-1/2} \sum_{j=1}^n X_j < x \right) - \Phi(x) \right| \leq c B_n^{-3/2} \sum_{j=1}^n \mathbf{E}\left[|X_j|^3 \right],$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of $\mathrm{N}(0,1)$.

Now, once we have enough employees so that no single new hire can have too dramatic an effect on the final gap after n hires, the conditional distribution of the final gap given the gap after these first few hires is approximately lognormal. In other words, the first few hires influence the distribution of the final gap a great deal, but once we condition on them, the distribution of the final gap is essentially lognormal. This idea is expressed formally in the following proposition.

Proposition 3.5. *Suppose $f : \mathbb{Z}_{>0} \to \mathbb{Z}_{>0}$ satisfies $f(n) = \omega(1)$ and $\limsup_{n\to\infty} f(n)/n < 1$. Then*

$$\sup_x \left| \mathbf{Pr}\left(\ln \frac{G_n}{G_{f(n)}} - \mathbf{E}\left[\ln \frac{G_n}{G_{f(n)}} \right] < x \, \mathbf{Var}\left[\ln \frac{G_n}{G_{f(n)}} \right] \right) - \Phi(x) \right| = O\left(f(n)^{-1/2} \right) = o(1),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of $\mathrm{N}(0,1)$.

Proof. Let $U_1, U_2, \ldots,$ be independent $\mathrm{Unif}(0,1)$ random variables, and define

$$Y_{n,i} = \ln\left(1 - \frac{U_i}{f(n)+i+1}\right) - \mathbf{E}\left[\ln\left(1 - \frac{U_i}{f(n)+i+1}\right) \right].$$

By Lemmas 3.1 and 3.2,

$$\ln \frac{G_n}{G_{f(n)}} - \mathbf{E}\left[\ln \frac{G_n}{G_{f(n)}} \right] \sim \sum_{i=1}^{n-f(n)} Y_{n,i}$$

and

$$\sum_{i=1}^{n-f(n)} \mathbf{Var}[Y_{n,i}] = \sum_{i=1}^{n-f(n)} \Omega\left(\frac{1}{(f(n)+i+1)^2} \right) = \Omega\left(f(n)^{-1} \right).$$

Also, for fixed n, the $Y_{n,i}$'s are independent, $\mathbf{E}[Y_{n,i}] = 0$, and

$$\begin{aligned}\sum_{i=1}^{n-f(n)} \mathbf{E}\left[|Y_{n,i}|^3\right] &= \sum_{i=1}^{n-f(n)} \mathbf{E}\left[\left(Y_{n,i}^4\right)^{3/4}\right] \\ &\leq \sum_{i=1}^{n-f(n)} \mathbf{E}\left[Y_{n,i}^4\right]^{3/4} \\ &= \sum_{i=1}^{n-f(n)} O\left(\frac{1}{(f(n)+i+1)^3}\right) \\ &= O\left(f(n)^{-2}\right),\end{aligned}$$

where the third step follows from Jensen's inequality and the fourth step follows from Lemma 3.2. Since

$$\Omega\left(f(n)^{-1}\right)^{-3/2} O\left(f(n)^{-2}\right) = O\left(f(n)^{-1/2}\right) = o(1),$$

an application of Lemma 3.3 completes the proof. □

Proposition 3.5 shows that, assuming we take our starting point to be after a sufficiently large number of hires, the body of the distribution of the final gap will be approximately lognormally distributed. The bounds given by Proposition 3.5, however, are weak at the tail of the distribution, since those bound give results only within $O\left(f(n)^{-1/2}\right) = O(n^{-1/2})$ (for, say, $f(n) = \lceil n/2 \rceil$). When we are dealing with subpolynomially small probabilities, such a bound is not useful.

To cope with this, we provide a martingale-based concentration bound, making use of the fact that changes in the average are generally small.

Proposition 3.6. *For any $s \geq 0$ and $t, \lambda > 0$, we have*

$$\begin{aligned}&\mathbf{Pr}\left(\left|G_{s+t} - G_s \prod_{j=1}^{t}\left(1 - \frac{1}{2(s+j)}\right)\right| \geq \lambda \;\middle|\; G_s\right) \\ &\leq 2\exp\left(-\frac{8\lambda^2(s+t+2)}{eG_s^2 \ln(1+t/(s+1))}\right).\end{aligned}$$

Proof. For $i = 0, \ldots, t$, let $X_i = \mathbf{E}[G_{s+t} \mid G_s, \ldots, G_{s+i}]$. Then the conditional distribution of the X_i's given G_s forms a martingale. For $i \geq 1$, let

$$M_i = \prod_{j=i}^{t}\left(1 - \frac{1}{2(s+j+1)}\right).$$

Then for $i \geq 1$

$$
\begin{aligned}
&|X_i - X_{i-1}| \\
&= |\mathbf{E}[G_{s+t} \mid G_s, \ldots, G_{s+i}] - \mathbf{E}[G_{s+t} \mid G_s, \ldots, G_{s+i-1}]| \\
&= |G_{s+i} M_{i+1} - G_{s+i-1} M_i| \\
&= \left| G_{s+i} - G_{s+i-1} \left(1 - \frac{1}{2(s+i+1)}\right) \right| M_{i+1} \\
&\sim \left| G_{s+i-1} \left(1 - \frac{\mathrm{Unif}(0,1)}{s+i+1}\right) - G_{s+i-1} \left(1 - \frac{1}{2(s+i+1)}\right) \right| M_{i+1} \\
&= \frac{G_{s+i-1}}{s+i+1} \left| \frac{1}{2} - \mathrm{Unif}(0,1) \right| M_{i+1} \\
&\leq \frac{G_s}{2(s+i+1)} M_{i+1} \\
&\leq \frac{G_s}{2(s+i+1)} \exp\left(-\frac{1}{2} \sum_{j=i+1}^{t} \frac{1}{s+j+1} \right) \\
&= \frac{G_s}{2(s+i+1)} \exp\left(-\frac{1}{2} \left[\sum_{j=1}^{s+t+1} \frac{1}{j} - \sum_{j=1}^{s+i+1} \frac{1}{j} \right] \right) \\
&\leq \frac{G_s}{2(s+i+1)} \exp\left(-\frac{1}{2} \left[\ln(s+t+2) - \ln(s+i+1) - 1 \right] \right) \\
&= \frac{G_s}{2} \sqrt{\frac{e}{(s+t+2)(s+i+1)}}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\sum_{i=1}^{t} |X_i - X_{i-1}|^2 &\leq \frac{G_s^2 e}{4(s+t+2)} \sum_{i=1}^{t} \frac{1}{s+i+1} \\
&\leq \frac{G_s^2 e}{4(s+t+2)} \int_{s+1}^{s+t+1} \frac{1}{x} \, \mathrm{d}x \\
&= \frac{G_s^2 e \ln(1 + t/(s+1))}{4(s+t+2)},
\end{aligned}
$$

and the result now follows immediately from Azuma's inequality. □

It is worth examining the bound of Proposition 3.6. When $s = 0$, $t = n$, so that the expected value of G_t is $\Theta(\sqrt{n})$, choosing values of λ that are $\Theta(1/\sqrt{n})$ gives useless bounds. This is not surprising given what we know regarding the distribution of G_n. However, if we choose $\lambda = (c \log n)/\sqrt{n}$ for some constant $c > 0$, we obtain inverse polynomial bounds on deviations from the expectation, and if we choose $\lambda = n^{-(1/2+\epsilon)}$ for any constant $\epsilon > 0$, we obtain probability bounds that are exponentially small.

4 Hiring Above the Median

We now analyze the hiring problem when the distribution of applicants' quality scores is $\text{Unif}(0,1)$ and one hires above the median. More precisely, we begin with one employee with quality $q \in (0,1)$. Whenever we have $2k+1$ employees, we hire the next two applicants with quality scores at least the median M_k of the $2k+1$ employees. The restriction that we only update the median when we have an odd number of employees is an interpretation that greatly simplifies the analysis.

One might suspect that hiring above the mean and hiring above the median would have essentially the same behavior, perhaps from the intuition that the median and means of several uniform random variables are generally quite close to each other. In fact this is not the case, as we now show. Hiring above the median leads to smaller gaps with fewer hires. Of course, the tradeoff is that the number of interviews between hires is much larger when hiring above the median than when hiring above the mean.

Our analysis is essentially analogous to that of Section 3. We begin with the following proposition, which we state without proof.

Proposition 4.1. *With probability 1, we hire infinitely many applicants and* $\lim_{k\to\infty} M_k = 1$.

We proceed by studying the *gap* sequence given by $G'_k \triangleq 1 - M_k$, which converges to 0 almost surely as $k \to \infty$. For convenience, we let $g = G'_0 = 1 - q$. Notice that G'_k refers to the setting where we have $2k+1$ employees.

4.1 Analysis of Expectations

Whereas in studying hiring above the mean we dealt with uniform distributions, when studying hiring above the median the beta distribution arises naturally. (Recall that $\text{Beta}(i,j)$ for integer values of i and j is the distribution of the ith smallest of a sample of $i+j-1$ independent $\text{Unif}(0,1)$ random variables; we use this fact repeatedly in our analysis.)

Lemma 4.1. *For any $t, k \geq 0$, the conditional distribution of G'_{t+k} given G'_k is the same as $G'_k \prod_{j=1}^{t} B_j$, where the B_j's are independent and $B_j \sim \text{Beta}(k+j+1, 1)$.*

Proof. The main idea of the proof is that when there are $2k+1$ employees, the quality scores of employees that are above M_k, as well as the quality scores of the next two hires, are independent $\text{Unif}(M_k, 1)$ random variables. M_{k+1} is therefore the minimum of $2k+3$ independent $\text{Unif}(M_k, 1)$ random variables, and the result follows by induction on k.

More formally, we proceed by induction on $t \geq 0$. For $t = 0$, the result is trivial. Now suppose that $t > 0$ and that the claim holds for $t-1$. Let $U_1, \ldots, U_{k+t+2}$ be independent $\text{Unif}(0,1)$ random variables, and let $U'_1, \ldots, U'_{k+t+1}$ be random variables that, given G'_{k+t-1}, are distributed as independent $\text{Unif}(G'_{k+t-1}, 1)$ random variables.

Now condition on $G' \triangleq (G_0, \ldots, G'_{k+t-1})$. Then the highest $k+t-1$ employee quality scores when there are $2(k+t-1)+1$ employees are distributed as $k+t-1$ independent $\text{Unif}(G'_{k+t-1}, 1)$ random variables. Furthermore, the quality scores of the next two hires are also independent $\text{Unif}(G'_{k+t-1}, 1)$ random variables. Thus, we have that M_{k+t} is distributed as

$$\begin{aligned}
\min\left\{U'_1, \ldots, U'_{t+k+1}\right\} &\sim \min\left\{1 - G'_{k+t-1}U_1, \ldots, 1 - G'_{k+t-1}U_{k+t+1}\right\} \\
&= 1 - G'_{k+t-1} \max\left\{U_1, \ldots, U_{k+t+1}\right\} \\
&\sim 1 - G'_{k+t-1}\text{Beta}(k+t+1, 1).
\end{aligned}$$

The conditional distribution of G'_{k+t} given G' is therefore given by $G'_{k+t-1}\text{Beta}(k+t+1,1)$, and the result now follows immediately from the induction hypothesis. □

With Lemma 4.1, we can find the expected gap $\mathbf{E}[G'_k]$, as well as the expected number of interviews to reach $2k+1$ employees, which turns out to have a very nice form. The proofs of the following propositions are analogous to those of Propositions 3.2 and 3.3.

Proposition 4.2.

$$\mathbf{E}[G'_k] = g\prod_{j=1}^{k}\left(1-\frac{1}{j+2}\right) = \Theta(1/k).$$

Proof. By Lemma 4.1,

$$\begin{aligned}\mathbf{E}[G'_k] = g\prod_{j=1}^{k}\mathbf{E}\left[\text{Beta}(j+1,1)\right] &= g\prod_{j=1}^{k}\left(1-\frac{1}{j+2}\right)\\ &= g\prod_{j=1}^{k}\exp\left(\ln\left(1-\frac{1}{j+2}\right)\right)\\ &= g\prod_{j=1}^{k}\exp\left(-1/j+\Theta\left(1/j^2\right)\right)\\ &= g\exp(-\ln k+\Theta(1)) = \Theta(1/k).\end{aligned}$$

□

Proposition 4.3. *Let T'_k be the number of interviews until there are $2k+1$ employees. Then $\mathbf{E}[T'_k] = k(k+1)/g$.*

Proof. For $i \geq 1$, let T''_i denote the number of interviews between when there are $2(i-1)+1$ employees and when there are $2i+1$ employees, so that $T'_k = \sum_{i=1}^{k} T''_i$. Then the conditional distribution of T''_i given G'_{i-1} is the sum of two independent geometric random variables with parameter G_{i-1}, and thus $\mathbf{E}[T''_i] = 2\,\mathbf{E}[1/G'_{i-1}]$. Now, by Lemma 4.1,

$$\mathbf{E}\left[\frac{1}{G'_{i-1}}\right] = \frac{1}{g}\prod_{j=1}^{i-1}\mathbf{E}\left[\frac{1}{\text{Beta}(j+1,1)}\right] = \frac{1}{g}\prod_{j=1}^{i-1}\frac{j+1}{j} = \frac{i}{g},$$

where the second step follows from integration. Thus,

$$\mathbf{E}[T'_k] = 2\sum_{i=1}^{k}\mathbf{E}\left[\frac{1}{G_{i-1}}\right] = \frac{2}{g}\sum_{i=1}^{k} i = \frac{k(k+1)}{g}.$$

□

Again, we note that after n^2 candidates, the expected value for the median gap for the best n candidates is $\Theta(1/n)$. Hence hiring above the median is, in this sense, within a constant factor of optimal.

Propositions 4.2 and 4.3 strongly suggest that the strategy of hiring above the median leads to higher quality employees than the strategy of hiring above the mean, at the cost of significantly slower company growth. However, one could reasonably believe that these results cannot be directly compared against those in Section 3.1. Indeed, in Section 3.1 we analyze the average quality score for the strategy of hiring above the mean, and in Proposition 4.2 we analyze the median quality score for the strategy of hiring above the median. Thus, we desire results about the median quality score for the strategy of hiring above the mean and the mean quality score for the strategy of hiring above the median. The former seems difficult, but we achieve the latter in Proposition 4.4 below.

Proposition 4.4. *Let A'_n denote the mean quality score of the first n employees when hiring above the median. Then $\mathbf{E}[A'_n] = 1 - \Theta(\log n/n)$.*

Proof. Let Q'_i denote the quality score of the ith hire. Then $Q'_0 = q$ and for $i \geq 1$, the conditional distribution of Q'_i given $G'_{\lceil \frac{i-2}{2} \rceil}$ is $\text{Unif}(1 - G'_{\lceil \frac{i-2}{2} \rceil}, 1)$. A simple calculation then gives

$$\mathbf{E}[Q'_i] = 1 - \frac{1}{2}\,\mathbf{E}\left[G'_{\lceil \frac{i-2}{2} \rceil}\right] = 1 - \Theta(1/i),$$

where we have used Proposition 4.2. It now follows that

$$\mathbf{E}[A'_n] = \frac{1}{n}\sum_{i=0}^{n-1} \mathbf{E}[Q'_i] = 1 - \frac{1}{n}\sum_{i=0}^{n-1} \Theta(1/i) = 1 - \Theta\left(\frac{\log n}{n}\right).$$

□

Proposition 4.4 can be directly compared with Proposition 3.2 to conclude that the strategy of hiring above the median really does give significantly higher quality employees than the strategy of hiring above the mean.

4.2 Convergence and Concentration

Hiring above the median also yields a weak convergence to a lognormal distribution, in the same sense as hiring above the mean. We can also obtain a martingale argument to handle the tails in this case, but the proof requires a more challenging argument. The additional difficulty comes from the fact that, for the strategy of hiring above the mean, a single new hire cannot change the mean employee quality scores significantly. However, when hiring above the median, a single new hire can have a drastic impact on the median employee quality score, although this is rather unlikely. Dealing with this difficulty makes the martingale argument slightly more complicated.

The proofs of the following two propositions are essentially analogous to those of Propositions 3.4 and 3.5. We begin with a technical lemma, which is easily checked by integrating appropriately.

Lemma 4.2. *For $j \geq 1$,*

1. $\mathbf{E}[\ln \text{Beta}(j+1,1)] = -\frac{1}{1+j}$
2. $\mathbf{Var}[\ln \text{Beta}(j+1,1)] = \frac{1}{(1+j)^2}$
3. $\mathbf{E}\left[(\ln \text{Beta}(j+1,1) - \mathbf{E}[\ln \text{Beta}(j+1,1)])^4\right] = \frac{9}{(1+j)^4}$

Proposition 4.5. *$\ln G'_k - \mathbf{E}[\ln G'_k]$ converges to some random variable G' almost surely as $k \to \infty$. Furthermore, the moment generating function of G' is $\psi_{G'}(t) = \Gamma(t+2)e^{(\gamma-1)t}$ for $t > -2$, where $\gamma = \lim_{k\to\infty} \sum_{j=1}^{k} \frac{1}{j} - \ln k$ is the Euler–Mascheroni constant.*

Proof. In light of Lemma 4.1, we may abuse the definition of our probability space and write $G'_k = g\prod_{j=1}^{k} B_j$, where the B_j's are independent and $B_j \sim \text{Beta}(j+1, 1)$. Letting $Y_j = \ln B_j - \mathbf{E}[\ln B_j]$ gives $Z_k \triangleq \ln G'_k - \mathbf{E}[\ln G'_k] = \sum_{j=1}^{k} Y_j$. Since $\mathbf{E}[Y_j] = 0$ and the Y_j's are independent, the Z_k's form a zero-mean martingale. Furthermore, by Lemma 4.2,

$$\mathbf{E}\left[Z_k^2\right] = \mathbf{Var}[Z_k] = \sum_{j=1}^{k} \mathbf{Var}[Y_j] = \sum_{j=1}^{k} \frac{1}{(1+j)^2} = O(1).$$

Therefore, we may apply a variant of the martingale convergence theorem (for example, [8, Theorem 7.8.1]) to obtain almost sure convergence of Z_k to some random variable G' as $k \to \infty$. Now, for $t > -2$, as $k \to \infty$,

$$\begin{aligned}
\mathbf{E}\left[e^{tZ_k}\right] &= \frac{\prod_{j=1}^{k} \mathbf{E}\left[\text{Beta}(j+1,1)^t\right]}{\exp\left(t\sum_{j=1}^{k} \mathbf{E}[\ln \text{Beta}(j+1,1)]\right)} \\
&= \frac{\prod_{j=1}^{k} \frac{j+1}{j+1+t}}{\exp\left(-t\sum_{j=1}^{k} \frac{1}{j+1}\right)} \\
&= \frac{\Gamma(t+2)\Gamma(k+2)}{\Gamma(k+2+t)} \exp\left(t\sum_{j=1}^{k} \frac{1}{j+1}\right) \\
&\sim \Gamma(t+2) \cdot \frac{(k+2)^{k+2}e^{-(k+2)}/\sqrt{k+2}}{(k+2+t)^{k+2+t}e^{-(k+2+t)}/\sqrt{k+2+t}} \\
&\qquad \times \exp\left(t\sum_{j=1}^{k} \frac{1}{j+1}\right) \\
&= \Gamma(t+2)\left(1 - \frac{t}{k+2+t}\right)^{k+2+t} \frac{e^t}{(k+2)^t}\sqrt{\frac{k+2+t}{k+2}} \\
&\qquad \times \exp\left(t\sum_{j=1}^{k} \frac{1}{j+1}\right) \\
&\sim \Gamma(t+2)e^{-t} \cdot \frac{e^t}{(k+2)^t} \cdot 1 \cdot \exp\left(t\sum_{j=1}^{k} \frac{1}{j+1}\right) \\
&\to \Gamma(t+2)e^{(\gamma-1)t} < \infty,
\end{aligned}$$

where the fourth step follows from Stirling's formula for the Gamma function. Since the Z_k's converge almost surely to G', we are done. □

Proposition 4.6. *Suppose $f : \mathbb{Z}_{>0} \to \mathbb{Z}_{>0}$ satisfies $f(k) = \omega(1)$ and $\limsup_{k\to\infty} f(k)/k < 1$. Then*

$$\sup_x \left| \mathbf{Pr}\left(\ln \frac{G'_k}{G'_{f(k)}} - \mathbf{E}\left[\ln \frac{G'_k}{G'_{f(k)}}\right] < x \,\mathbf{Var}\left[\ln \frac{G'_k}{G'_{f(k)}}\right] \right) - \Phi(x) \right| = O\left(f(k)^{-1/2}\right) = o(1),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of $\mathrm{N}(0,1)$.

Proof. Let $B_1, B_2, \ldots$ be independent random variables with $B_j \sim \mathrm{Beta}(j+2, 1)$ and let $Y_{k,j} = \ln B_{f(k)+j} - \mathbf{E}[\ln B_{f(k)+j}]$. By Lemmas 4.1 and 4.2,

$$\ln \frac{G'_k}{G'_{f(k)}} - \mathbf{E}\left[\ln \frac{G'_k}{G'_{f(k)}}\right] \sim \sum_{j=1}^{k-f(k)} Y_{k,j}$$

and

$$\sum_{j=1}^{k-f(k)} \mathbf{Var}[Y_{k,j}] = \sum_{j=1}^{k-f(k)} \Omega\left((f(k)+j)^{-2}\right) = \Omega\left(f(k)^{-1}\right).$$

Also, for fixed k, the $Y_{k,j}$'s are independent, $\mathbf{E}[Y_{k,j}] = 0$, and

$$\begin{aligned}
\sum_{j=1}^{k-f(k)} \mathbf{E}\left[|Y_{k,j}|^3\right] &= \sum_{j=1}^{k-f(k)} \mathbf{E}\left[\left(Y_{k,j}^4\right)^{3/4}\right] \\
&\le \sum_{j=1}^{k-f(k)} \mathbf{E}\left[Y_{k,j}^4\right]^{3/4} \\
&= \sum_{j=1}^{k-f(k)} O\left((f(k)+j)^{-3}\right) \\
&= O\left(f(k)^{-2}\right),
\end{aligned}$$

where the third step follows from Jensen's inequality and the fourth step follows from Lemma 4.2. Since

$$\Omega\left(f(k)^{-1}\right)^{-3/2} O\left(f(k)^{-2}\right) = O\left(f(k)^{-1/2}\right) = o(1),$$

an application of Lemma 3.3 completes the proof. □

As before, we may use a martingale argument to obtain bounds where Proposition 4.6 breaks down, although the argument is a bit more sensitive.

Proposition 4.7. *For $s, t \ge 1$, $u \le s$, and $\lambda > 0$, we have*

$$\begin{aligned}
&\mathbf{Pr}\left(\left| G'_{s+t} - G'_s \prod_{j=1}^{t} \left(1 - \frac{1}{s+j+2}\right) \right| \ge \lambda \;\middle|\; G_s \right) \\
&\le 2\exp\left[-\frac{\lambda^2}{2t} \left(\frac{s+t+2}{euG'_s} \right)^2 \right] + te^{-u+1}.
\end{aligned}$$

Proof. For $i = 0, \ldots, t$, let $X_i = \mathbf{E}[G'_{s+t} \mid G'_s, \ldots, G'_{s+i}]$. Then the conditional distribution of the X_i's given G'_s forms a martingale, to which we wish to apply some variant of Azuma's inequality. Unfortunately, this requires an effective bound for $|X_i - X_{i-1}|$ which is difficult, as there is some chance that when the median changes there is an unusually large difference between the corresponding gaps. While these events are rare, we must take them into account.

Let us suppose we have values c_i and b_i such that $|X_i - X_{i-1}| \le c_i$ with probability $1 - b_i$. Then it follows (in a manner similar to [3, Theorem 5.1]) that we can use the following Azuma's inequality variant:

$$\mathbf{Pr}\left(|X_t - X_0| > \lambda\right) \le 2e^{-\lambda^2/2\sum_{i=1}^{t} c_i^2} + \sum_{i=1}^{n} b_i.$$

Intuitively, we think of having a bad event when $|X_i - X_{i-1}| > c_i$, and separate out the probability of any bad event; if no bad events occur, we have a well-behaved martingale.

Now, we know that the distribution of G'_{s+i} given G'_{s+i-1} has the form $G'_{s+i-1}Z_i$, where Z_i is a random variable with distribution $\mathrm{Beta}(s+i+1, 1)$, which has expectation $1-1/(s+i+2)$. Let us consider the event $Z_i \le 1-u/(s+i+2) \triangleq z_i$, which occurs with probability $(1-u/(s+i+2))^{s+i+1} \le e^{-u+1}$. Now for $i \ge 1$, if $Z_i > z_i$,

$$\begin{aligned}
|X_i - X_{i-1}| &= \left|\mathbf{E}[G'_{s+t} \mid G'_s, \ldots, G'_{s+i}] - \mathbf{E}[G'_{s+t} \mid G'_s, \ldots, G'_{s+i-1}]\right| \\
&= \left|G'_{s+i}\prod_{j=i+1}^{t}\left(1-\frac{1}{s+j+2}\right) - G'_{s+i-1}\prod_{j=i}^{t}\left(1-\frac{1}{s+j+2}\right)\right| \\
&= \left|G'_{s+i} - G'_{s+i-1}\left(1-\frac{1}{s+i+2}\right)\right|\prod_{j=i+1}^{t}\left(1-\frac{1}{s+j+2}\right) \\
&\sim \left|G'_{s+i-1}Z_i - G'_{s+i-1}\left(1-\frac{1}{s+i+2}\right)\right|\prod_{j=i+1}^{t}\left(1-\frac{1}{s+j+2}\right) \\
&= G'_{s+i-1}\left|1-\frac{1}{s+i+2} - Z_i\right|\prod_{j=i+1}^{t}\left(1-\frac{1}{s+j+2}\right) \\
&\le G'_s\frac{u}{s+i+2}\prod_{j=i+1}^{t}\left(1-\frac{1}{s+j+2}\right) \\
&\le \frac{uG'_s}{s+i+2}\exp\left(-\sum_{j=i+1}^{t}\frac{1}{s+j+2}\right) \\
&= \frac{uG'_s}{s+i+2}\exp\left(-\left[\sum_{j=1}^{s+t+2}\frac{1}{j} - \sum_{j=1}^{s+i+2}\frac{1}{j}\right]\right) \\
&\le \frac{uG'_s}{s+i+2}\exp\left(-\left[\ln(s+t+2) - \ln(s+i+2) - 1\right]\right) \\
&= \frac{euG'_s}{s+t+2},
\end{aligned}$$

	Hiring above the Mean		Hiring above the Median	
Hires	Average Gap	Expected Gap	Average Gap	Expected Gap
256	0.03556	0.03518	0.00761	0.00769
512	0.02493	0.02490	0.00370	0.00388
1024	0.01755	0.01762	0.00189	0.00195
2048	0.01251	0.01246	9.68×10^{-4}	9.75×10^{-4}
4096	0.00867	0.00881	4.89×10^{-4}	4.88×10^{-4}
8192	0.00630	0.00623	2.47×10^{-4}	2.44×10^{-4}
16384	0.00444	0.00441	1.25×10^{-4}	1.22×10^{-4}
32768	0.00313	0.00312	6.09×10^{-5}	6.10×10^{-5}
65536	0.00223	0.00220	3.10×10^{-5}	3.05×10^{-5}

Table 1: Average gap values (from simulation) and their expected values (calculated).

and so

$$\sum_{i=1}^{t} |X_i - X_{i-1}|^2 \leq t \left(\frac{euG'_s}{s+t+2} \right)^2.$$

The result now follows directly from the variant of Azuma's inequality given above. □

Proposition 4.7 tells essentially the same story as Proposition 3.6. For $s = 1$ and $t = n$, we have $\mathbf{E}[G_t] = \Theta(1/n)$, and so choosing values of λ that are $\Theta(1/n)$ gives useless bounds. As before, this is not surprising. However, if we fix $s = u = \lceil c_1 \ln n \rceil$, $G'_s = \Theta(\mathbf{E}[G'_s]) = \Theta(1/s)$, and $\lambda = \sqrt{(c_2 \ln n)/n}$ for some constants $c_1, c_2 > 1$, then we obtain inverse polynomial bounds on deviations from the expectation. Similarly, if we choose $s = u = \lceil c(\ln n)^2 \rceil$, $G'_s = \Theta(\mathbf{E}[G'_s]) = \Theta(1/s)$, and $\lambda = n^{-(1/2+\epsilon)}$ for some constants $c > 1$ and $\epsilon > 0$, we obtain probability bounds that are sub-polynomially small.

5 Simulations

In this section, we present simulation results related to our analysis. We give these results with two goals in mind. First, we wish to check our theoretical analysis of the expected values associated with these processes against simulation results. Second, we wish to verify and examine the relationship between the gap distributions and the lognormal distributions, with a particular emphasis on the tails.

In Tables 1 and 2, we provide the average (that is, the mean) and expected number of interviews and gaps from our simulations, each of which consists of 1000 trials, starting with one employee with quality score 0.5. (The expectations are calculated directly from the formulas in Propositions 3.2, 3.3, 4.2, and 4.3.) As can be seen, across the board the simulation numbers match the exact answers obtained from our analysis.

Examining the lognormal approximation discussed in Sections 3.2 and 4.2 through simulation requires a little more work. The results in those sections tell us that, for either the strategy of hiring above the mean or the strategy of hiring above the median, if we let H_n denote the threshold gap score for hiring a new employee after hiring n people, then for any $k \ll n$, we have that

$$Z_{n,k} \triangleq \frac{\ln H_n - \mathbf{E}[\ln H_n \mid H_k]}{\mathbf{Var}[\ln H_n \mid H_k]} \approx \mathrm{N}(0,1),$$

	Hiring above the Mean		Hiring above the Median	
Hires	Average Interviews	Expected Interviews	Average Interviews	Expected Interviews
256	5192	5261	33560	33024
512	14865	14856	133547	131584
1024	41987	41986	561442	525312
2048	118011	118706	2408235	2099200
4096	340552	335682	8195709	8392704
8192	939643	949352	35427404	33562624
16384	2669107	2685031	130726647	134234112
32768	7585301	7594213	548812815	536903680
65536	21273638	21479393	2180257224	2147549184

Table 2: Average numbers of interviews (from simulation) and their expected values (calculated).

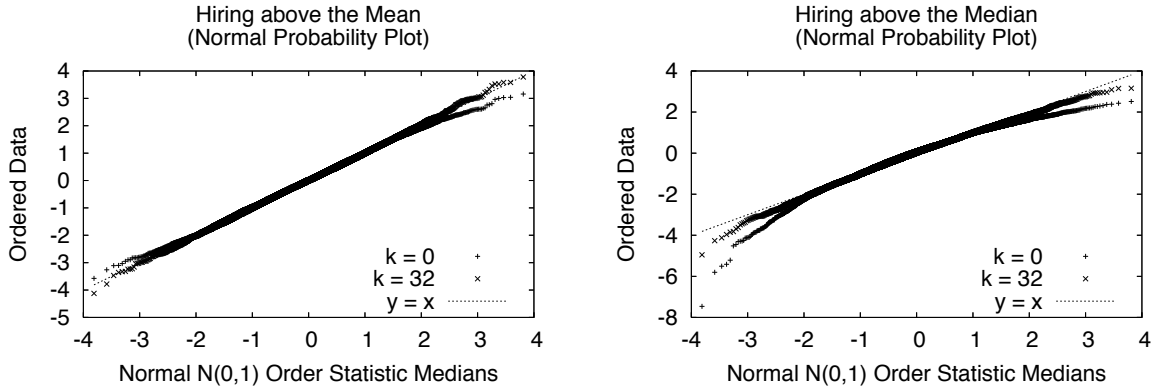

Figure 1: Plots of the samples of $Z_{1024,k}$.

where the approximation is in the sense of probability distributions. Furthermore, this approximation is good for the body of the distribution of $Z_{n,k}$, but fairly inaccurate at the tails.

To demonstrate these claims via simulation, we set $n = 1024$ and the quality score of the first employee to 0.5. Then we take 10000 independent samples of $Z_{n,k}$ for $k = 0$ and $k = 32$. (Note that the conditional expectations and variances needed to compute $Z_{n,k}$ from H_n and H_k are easily determined from the results in Sections 3 and 4.) We graph the results in Figure 1 using normal probability plots [13]. (Intuitively, each graph is obtained by plotting the samples so that, if they were truly drawn from $\mathrm{N}(0, 1)$, they would all lie very close to the line $y = x$ with high probability.)

It is clear from Figure 1 that the approximation $Z_{n,k} \approx \mathrm{N}(0, 1)$ is fairly accurate for the body of the distribution of $Z_{n,k}$, but weaker at the tails. It is also evident that as k increases, the approximation improves, which tells us that first few hires really do have a substantial effect for both strategies. Furthermore, we see that the normal approximation is better for the strategy of hiring above the mean than for the strategy of hiring above the median, which indicates that the latter strategy is more sensitive to the quality scores of the first few hires.

By using standard techniques for interpreting normal probability plots [13], we can also see how the tails of $Z_{n,k}$ differ from those of $\mathrm{N}(0, 1)$. Indeed, both curves on the graph for hiring above the median depart markedly downwards from the line $y = x$ on the tails, especially on the lower tail, which tells us that, in this case, the distribution of $Z_{n,k}$ has a long left tail and a short right tail, and that this effect diminishes as k grows. Since lower values of $Z_{n,k}$ correspond to higher employee quality scores, this observation tells us that for the strategy of hiring above the median, the first few hires really do have a tremendous impact on the final result, and this impact is much more likely to be positive than negative.

6 Variations and Extensions

One of the key features of the hiring model is that it naturally allows for variations and extensions, which may be useful for considering more realistic scenarios or gaining more insight into the underlying tradeoffs. While we expect other researchers will consider more elaborate extensions in future work, we briefly consider some natural extensions here, focusing on situations where our analysis can be generalized easily.

6.1 Preprocessing Interviews

The number of interviews required to hire an employee for the Lake Wobegon strategies starts small but grows quickly. As interviews are themselves expensive, the fact that the number of interviews grows in this way suggests a potential problem with our model.

In reality we expect interview preprocessing to occur. This preprocessing may simply stem from self-selection; low quality people do not bother to apply. Alternatively, a weeding process could discard weak candidates early in the process, such as after a read of the résumé instead of after a full interview.

Such preprocessing does not substantially affect our model, as long as the conditional distribution of the quality of a person above the current hiring threshold remains uniform. That is, the quality of the ith hire does not need to be different, just the number of interviews to reach the person, which can be handled separately. As an example, if when hiring above the mean the applicants have quality uniform over $(1 - cG_t, 1)$ for some constant $c > 1$ and current gap G_t, then

on average only c interviews are needed per hire, but the change in the gap over each hire follows the same distribution. One can devise similar models for hiring above the median.

6.2 Alternative Quantile Hiring Strategies

In some sense, there is really nothing special about the median in the hiring problem; we could consider hiring people above the 60th percentile of the current work force instead, for example.

For purposes of analysis, the easiest way to generalize our previous work is to consider strategies of the following type: fix some employee Q in the ranking, hire a additional people whose quality is above Q, and then move the threshold up b employees in the rank order. Our median strategy can be thought of as the "hire 2, move up 1" strategy, and we generalize this to a "hire a, move up b" strategy. Other quantiles can be suitably approximated by an appropriate choice of a and b, and the analysis in Section 4 can be appropriately generalized as follows. Let G_k^* denote the quality at the threshold employee after ka hires. Then the proof of Lemma 4.1 (corresponding to the "hire 2, move up 1" strategy) can be immediately generalized.

Lemma 6.1. *For any $t, k \geq 0$, the conditional distribution of G_{t+k}^* given G_k^* is the same as $G_k^* \prod_{j=1}^{t} B_j$, where the B_j's are independent and $B_j \sim \text{Beta}((a-b)(k+j)+1, b)$.*

Following Section 4, we use Lemma 6.1 to calculate various important quantities related to this hiring strategy. We therefore have in this setting, starting with $G_0^* = g$,

$$\begin{aligned}
\mathbf{E}[G_k^*] &= g \prod_{j=1}^{k} \mathbf{E}\left[\text{Beta}((a-b)j+1, b)\right] \\
&= g \prod_{j=1}^{k} \frac{(a-b)j+1}{(a-b)j+1+b} \\
&= g \prod_{j=1}^{k} \left(1 - \frac{b}{(a-b)j+1}\right) \\
&= g \prod_{j=1}^{k} \exp\left(\ln\left(1 - \frac{b}{(a-b)j+1}\right)\right) \\
&= g \prod_{j=1}^{k} \exp\left(-b/(a-b)j + \Theta\left(1/j^2\right)\right) \\
&= g \exp(-b/(a-b)\ln k + \Theta(1)) = \Theta(k^{-b/(a-b)}).
\end{aligned}$$

For $i \geq 1$, let T_i^{**} denote the number of interviews between when there are $a(i-1)+1$ employees and when there are $ai+1$ employees, so that $T_k^* = \sum_{i=1}^{k} T_i^{**}$ is the total number of interviews before there are $ak+1$ employees. Then the conditional distribution of T_i^{**} given G_{i-1}^* is the sum of a independent geometric random variables with parameter G_{i-1}^*, and thus $\mathbf{E}[T_i^{**}] = a\,\mathbf{E}[1/G_{i-1}^*]$.

Now, by Lemma 6.1,

$$\begin{aligned}\mathbf{E}\left[\frac{1}{G^*_{i-1}}\right] &= \frac{1}{g}\prod_{j=1}^{i-1}\mathbf{E}\left[\frac{1}{\text{Beta}((a-b)j+1,b))}\right] \\ &= \frac{1}{g}\prod_{j=1}^{i-1}\frac{(a-b)j+b}{(a-b)j} \\ &= \Theta(i^{b/(a-b)}),\end{aligned}$$

where the second step follows from integration. Thus,

$$\mathbf{E}[T^*_k] = a\sum_{i=1}^{k}\mathbf{E}\left[\frac{1}{G'_{i-1}}\right] = \Theta(k^{a/(a-b)}).$$

Similarly to the strategies of hiring above the mean and the median, here we have that after $n^{a/(a-b)}$ candidates, the expected value of the $\frac{b}{a}n$ order statistic of the smallest n gap scores is $\Theta(n^{-b/(a-b)})$. Hence the "hire a, move up b" strategies are, in this sense, within a constant factor of optimal.

6.3 Errors

In our analyses in Sections 3 and 4, we assume that an applicant's exact quality score is revealed during the interview. In reality, however, the interview process cannot be perfect, so the interview score will differ from the applicant's true quality. We may therefore hire applicants whose true quality scores lie below the threshold prescribed by the hiring strategy, and similarly we may reject applicants whose scores lie above this threshold. Furthermore, it may also be unrealistic to assume that we know the exact value of the threshold when interviewing applicants, as this may require more information about our employees' quality scores than we can exactly determine.

We would like to take these sorts of errors into account in our analysis. Unfortunately, it seems quite difficult to formulate a model for all, or even just some, of these errors that is simultaneously justifiable and analyzable. In particular, our analysis in Sections 3 and 4 relies heavily on the fact that the conditional distribution of an applicant's quality score Q_i given that $Q_i \geq x$ is $\text{Unif}(x,1) \sim 1-(1-x)\,\text{Unif}(0,1)$. This observation allows us to derive the simple expressions for the form of the gap distribution given in Lemmas 3.1 and 4.1. But if we allow for errors in the interview process, then the conditional distribution of an applicant's true quality score Q_i given that the observed score $\hat{Q}_i \geq x$ does not seem to have an analogous form for most standard models of measurement error. We believe that resolving this issue is a worthwhile open problem.

As an example of what we can analyze, for the case of hiring above the mean, suppose that the conditional distribution of the true quality score of the ith employee hired given the prior history of the system and $G_{i-1} \leq g_0$ is $\text{Unif}(1-G_{i-1}R_{i-1},1)$, where G_{i-1} is the true gap and the R_i's are random variables with common distribution $0 < R \leq 1/g_0$ that are independent of each other and the applicants' true quality scores; the case $R = 1$ corresponds to the model analyzed in Section 3. This model is somewhat artificial, but it captures the idea that there may be noise in our observations of our quality scores.

Going through the same calculations as in the proof of Lemma 3.1 now tells us that, given $G_s \leq g_0$, the conditional joint distribution of the G_i's for $i > s$ is the same as if $G_i = G_s\prod_{j=s+1}^{i}\left(1-\frac{1-R_jU_j}{j}\right)$,

where the U_j's are independent Unif$(0,1)$ random variables that are also independent of the R_j's. This formula could be used to prove analogues of the results in Section 3.

6.4 Hiring and Firing

Another natural extension to the hiring problem would be to allow a firing strategy, in order to remove the low performers.[2] Intuitively, a good firing strategy should allow the company to optimize the tradeoff between increasing its employees' average quality and reducing their number. Unfortunately, most natural combinations of hiring and firing strategies seem difficult to analyze, because they introduce challenging dependencies among employee quality scores. We can, however, partially analyze an important class of firing strategies in conjunction with hiring above the median. These strategies are generalizations of the (in)famous *rank-and-yank* system (sometimes, and perhaps more properly, called the *vitality curve* system), used extensively by Jack Welch at General Electric [18].

The basic tenet of the rank-and-yank system is that employees should be periodically ranked and those near the bottom should be fired. The key (and most controversial) detail of this system is that the fraction of employees fired is fixed in advance, without regard to any absolute or overall measurements of the employees' qualities or performance. Such a strategy is easily modeled in the context of hiring above the median examined in Section 4, assuming that quality scores do not change over time. Indeed, if candidates' quality scores are independent Unif$(0,1)$ random variables and we condition on having $2k+1$ employees with median quality score M at some particular time, then the distribution of the top k employees' quality scores is the same as the order statistics of k independent Unif$(M,1)$ random variables. Thus, if we were to fire the bottom $2j \leq 2k$ employees at this time, the conditional distribution of the resulting median of the employees' quality scores would be the same as the distribution of the jth smallest of k independent Unif$(M,1)$ random variables, which is $1-(1-M)\text{Beta}(k-j,j+1)$.

With this in mind, we consider the following variant of hiring above the median in conjunction rank-and-yank firing. We start with one employee with some fixed quality $0 < q < 1$, and we fix some *firing parameter* $0 < f < 1$. Whenever we have $2k+1$ employees, we first hire the next two applicants whose quality scores are above the current median of the current employees' quality scores, giving a total of $2k+3$ employees. If $(2k+2)f$ is an even integer, we then fire the $(2k+2)f$ employees with lowest quality scores. (Note that the number of employees is always odd, so there is no ambiguity in determining the median. Also, if $f = 0$, the model is exactly the same as the one studied in Section 4.)

Let G_t'' and $n(t)$ denote the gap and number of employees after t iterations of this process, so that $G_0'' = g = 1-q$ and $n(0) = 1$. For convenience, let $m(t) = (n(t)-1)/2$ and let $r(t)$ denote the (deterministic) number of employees fired during the t-th iteration. Then by the argument above, we immediately have a natural analogue to Lemma 4.1.

Lemma 6.2. *For $t \geq 1$, the conditional distribution of G_t'' given everything that occurs in the first $t-1$ iterations of the process is the same as $G_{t-1}''\text{Beta}(m(t-1)+2-r(t), r(t)+1)$.*

This lemma provides a starting point from which more detailed analyses of rank-and-yank strategies could proceed.

[2] It is interesting to note that under our simple model, the lowest performers are always the ones with the most seniority. Hence there may be some truth to the essentially universally held belief that you are better than your boss.

6.5 Quality Changes

A direction we have yet to consider, but that seems ripe for future work, would be to consider how the hiring model behaves if candidate quality scores are themselves variable over time. In this way, early hires could improve and become more valuable to the company over time, and candidates who appear strong at their initial interview could eventually become ripe for firing. Introducing time-varying dynamics would require significantly more insight into the structure of the work environment, taking us fairly far afield from the more universal starting point of the original hiring problem. However, we hope that the basic insights from the simple hiring problem may prove useful for more detailed economic analyses of hiring and firing dynamics.

7 Conclusions

We have introduced the hiring problem, a mathematical model for decision making under uncertainty related to the secretary problem. We have also introduced and analyzed the behavior of Lake Wobegon strategies, where one hires new applicants that lie above the mean or the median of the current employees. These simple scenarios already provide rich mathematical structures, with connections to lognormal distributions, weak convergence results, and nonintuitive differences between the mean and median. Furthermore, the large number of possible variations and extensions suggests that there are many more interesting connections and developments yet to make.

References

[1] M. Babaioff, N. Immorlica, and R. Kleinberg. Matroids, secretary problems, and online mechanisms. In *Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms*, pages 434–443, 2007.

[2] A. Z. Broder, A. Kirsch, R. Kumar, M. Mitzenmacher, E. Upfal, and S. Vassilvitskii. The hiring problem and Lake Wobegon strategies. In *Proceedings of the 19th ACM-SIAM Symposium on Discrete Algorithms*, 2008 (to appear). Available at `http://www.eecs.harvard.edu/~kirsch/pubs/hiring/soda.pdf`.

[3] F. Chung and L. Lu. Coupling Online and Offline Analyses for Random Power Law Graphs. *Internet Mathematics*, 1(4):409–461, 2004.

[4] T. Ferguson. Who solved the secretary problem? *Statistical Science*, 4(3):282–296, 1989.

[5] P. R. Freeman. The secretary problem and its extensions: A review. *International Statistical Review*, 51(2):189–206, 1983.

[6] M. Gardner. Mathematical games. *Scientific American*, pages 150–153, 1960.

[7] K. S. Glasser, R. Holzsager, and A. Barron. The d choice secretary problem. *Communications in Statistics C-Sequential Analysis*, 2:177–199, 1983.

[8] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, third edition, 2003.

[9] M. T. Hajiaghayi, R. Kleinberg, and D. C. Parkes. Adaptive limited-supply online auctions. In *Proceedings of the 5th ACM Conference on Electronic Commerce*, pages 71–80, 2004.

[10] N. Immorlica, R. Kleinberg, and M. Mahdian. Secretary problems with competing employers. In *Proceedings of the 2nd Workshop on Internet and Network Economics*, pages 389–400, 2006.

[11] R. Kleinberg. A multiple-choice secretary problem with applications to online auctions. In *Proceedings of the 16th ACM-SIAM Symposium on Discrete Algorithms*, pages 630–631, 2005.

[12] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[13] NIST/SEMATECH e-Handbook of Statistical Methods. Normal Probability Plot, `http://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm`, 2006. Accessed November 30, 2006.

[14] P. Norvig. Hiring: The Lake Wobegon strategy, March 11, 2006. Google Research Blog, `http://googleresearch.blogspot.com/2006/03/hiring-lake-wobegon-strategy.html`.

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

[16] S. Samuels. Secretary problems. In *Handbook of Sequential Analysis*, pages 381–405. Marcel Dekker, Inc., 1991.

[17] R. J. Vanderbei. The optimal choice of a subset of a population. *Mathematics of Operations Research*, 5:481–486, 1980.

[18] Wikipedia. Vitality curve, `http://en.wikipedia.org/w/index.php?title=Vitality_curve&oldid=91009444`, 2006. Accessed November 30, 2006.

[19] J. G. Wilson. Optimal choice and assignment of the best m of n randomly arriving items. *Stochastic Processes and their Applications*, 39(2):325–343, 1991.