Crowding at the front of marathon packs

J. Stat. Mech. (2008) L03001

(http://iopscience.iop.org/1742-5468/2008/03/L03001)

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

LETTER

Crowding at the front of marathon packs

Sanjib Sabhapandit[1], **Satya N Majumdar**[1] **and S Redner**[2]

[1] Laboratoire de Physique Théorique et Modèles Statistiques (UMR 8626 du CNRS), Université Paris-Sud, Bâtiment 100, 91405 Orsay Cedex, France
[2] Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, USA
E-mail: sanjib.sabhapandit@u-psud.fr, majumdar@lptms.u-psud.fr and redner@bu.edu

Received 13 February 2008
Accepted 3 March 2008
Published 18 March 2008

Online at stacks.iop.org/JSTAT/2008/L03001
doi:10.1088/1742-5468/2008/03/L03001

Abstract. We study the crowding of near-extreme events in the time gaps between successive finishers in major international marathons. Naively, one might expect these gaps to become progressively larger for better-placing finishers. While such an increase does indeed occur from the middle of the finishing pack down to approximately 20th place, the gaps saturate for the first 10–20 finishers. We give a probabilistic account of this feature. However, the data suggest that the gaps have a weak maximum around the 10th place, a feature that seems to have a sociological origin.

Keywords: stochastic processes, traffic and crowd dynamics

ArXiv ePrint: 0802.1702

J. Stat. Mech. (2008) L03001

It is fun to learn about sports statistics and discuss their implications among fellow sports fans. The existence of comprehensive Web-based resources for sports statistics, whose easy availability was unimaginable just a few years ago, has perhaps helped promote such activities. In this note, we investigate one such statistic, namely, the finishing times of individual runners in major marathons [1]. Our main interest is in the dependence of the *time gaps between successive finishers* on finishing place. More precisely, let t_k be the time of the kth finisher. Then we wish to understand how the time gaps $g_k \equiv t_{k+1} - t_k$ depend on finishing place k. Because front runners are rare and potential race leaders are rarer still, the natural expectation is that the gaps between successive finishers should increase monotonically in moving from the middle of the pack towards the increasingly rare front runners. However, the data show that the time gaps saturate to a constant value for sufficiently small k. We suggest that sociological factors may contribute to this anomaly in the gaps.

The results presented here are based on data for finishing times in major international marathons that attract world-class entrants. These include Boston, Chicago, and New York from 2000 to 2007 (entire fields), as well as Berlin 1992 and 1999–2007, Fukuoka 2006–2007, London 2001–2007, and Paris 2004, 2006–2007 (first 100 places for all non-US races). Data for other years in these non-US marathons are not readily available or corrupted, and some of the data used in this work required corrections of a few obviously erroneous results. In these marathons, the winning time is in the range 2:05–2:10. For example, in Boston, Chicago, and New York, the course records are 2:07:14, 2:05:42, and 2:07:43, respectively, while the current world record, set by Haile Gebrselassie in the 2007 Berlin Marathon, is 2:04:26. After the race winner, there is a trickle of fast finishers that gradually turns into a steady flow as the finish time approaches 3 h. The main pack arrives in the range of 3–6 h, with a decreasing stream of progressively slower stragglers. Thus one naturally anticipates the distribution of finish times shown in figure 1.

Upon examining these distributions critically, a number of curiosities can be seen. First, in spite of the data smoothing, there are visible peaks at just under 3 and 4 h for all three marathons. For the Chicago marathon in particular, where the course is flat and well suited for pacing, one can even discern secondary peaks near 3:10, 3:20, and 3:30 (figure 1). The existence of such peaks suggests that the distribution of finish times in this range does not reflect a performance limit, but rather, the surmounting of a psychological barrier. Parenthetically, the apparent difference in the distributions for the Boston marathon (where challenging qualifying times exist), and the Chicago and New York marathons can be made to nearly disappear by plotting them in scaled units—namely, by making the abscissa the finish time divided by the average finish time for each set of eight races.

More interesting behavior, and the main point of this work, is the k dependence of the time gaps g_k between successive finishers. We are particularly interested in these gaps for finishers near the front of the pack. Thus we restrict ourselves to the first 10 000 finishers in the US marathons. This threshold corresponds to finishing times of about 4 h for Chicago and New York, and around 3:45 for Boston. By comparing with figure 1, these time thresholds are prior to the peak of the finishing time distribution for Chicago and New York, and near the peak for Boston. For comparison, the average number of finishers over the last eight US marathons that we studied is 30 668 for Chicago, 33 669 for New York, and 16 645 for Boston. For $k > 10\,000$, $\langle g_k \rangle$ begins increasing, corresponding

doi:10.1088/1742-5468/2008/03/L03001

J. Stat. Mech. (2008) L03001

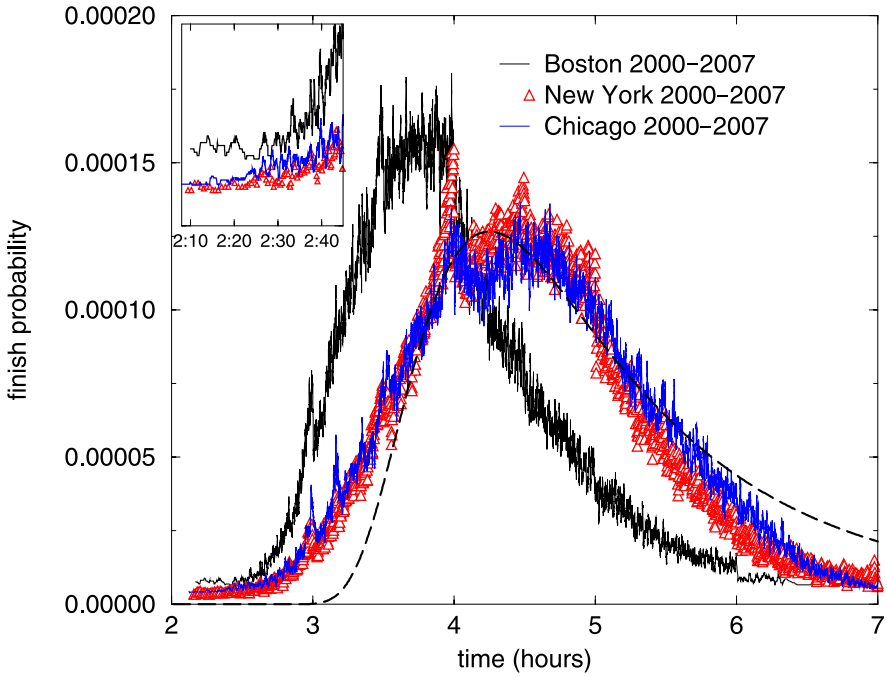

Figure 1. Distribution of all finishing times (smoothed over a 20-point range for visual clarity) for the Boston, Chicago, and New York marathons, 2000–2007. Notice the peaks at 3 h in all the data, the prominent peaks at 4 h for the Chicago and New York marathons, and the secondary peaks at 3:10, 3:20, and 3:30 for the Chicago marathon. The dashed curve shows the distribution of equation (4), with parameter values as given in the text. The inset shows the data in the range of 2:08–2:45.

to the lagging tail of the finishing time distribution. For the European data, we quote $g(k)$ only to $k = 100$.

Among the fastest finishers, the finish time distribution decays very slowly and is nearly constant for times less than 2:30 (inset to figure 1). For the marathons that we studied, the *average* time gap between consecutive finishers among the first 10 places is in the range of 20–60 s, and does not have any clear systematic k dependence (figure 2). Members of this group of elite runners are all possible candidate winners of the race on any given day. In contrast, beyond the 20th place, the average gap systematically decreases with k, a decrease that clearly reflects the increase in the density of runners as the leading edge of the pack arrives at the finish.

We can make these observations quantitative by assuming that the finishing times of individual runners are independent and identically distributed (iid) random variables, and then using extreme-value statistics to determine the time gaps g_k between successive finishers [2, 3]. As a preliminary, consider the time t_k of the kth finisher. The *typical* value for this time can be determined from the extremal condition (which assumes self-averaging)

$$\int_0^{t_k} P(t)\,\mathrm{d}t \approx \frac{k}{N}, \tag{1}$$

J. Stat. Mech. (2008) L03001

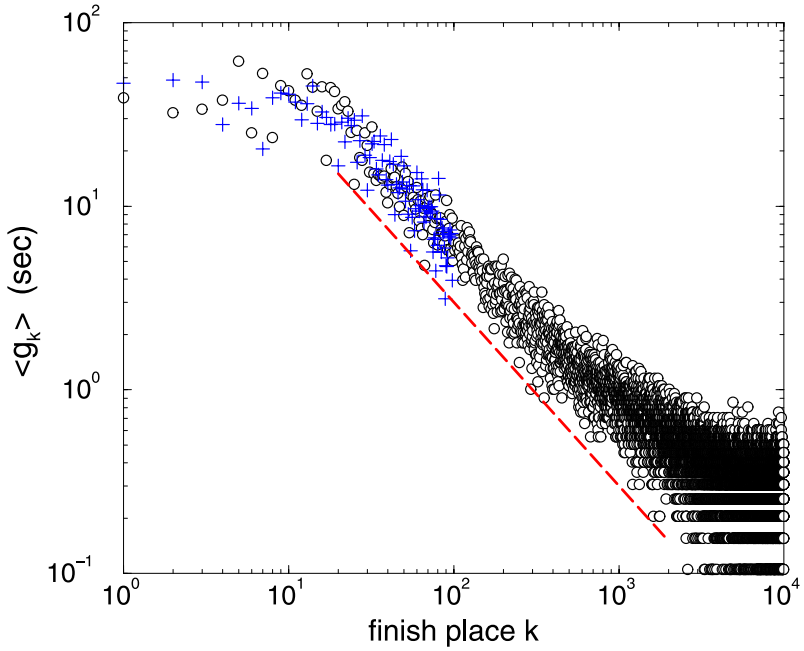

Figure 2. Distribution of the average time gap g_k between the kth and $(k+1)$st finisher for: the US (○) and the European (+) marathons cited in the text. For the US marathons the first 10 000 gaps are shown, while the first 100 gaps are shown for the European marathons. The dashed line has a slope of -1, as given by equation (5).

that states that there are k individuals whose finishing times are less than t_k. The resulting estimate for the typical kth finishing time t_k should be accurate for $k \gg 1$, where fluctuations in t_k are negligible. More generally, we can compute the full probability distribution of t_k, as outlined in appendix A, and thereby find the mean value of t_k to be

$$\langle t_k \rangle = \int_0^\infty I\left(P_>(x); N-k+1, k\right)\, \mathrm{d}x, \tag{2}$$

where $I(y; a, b) = [\int_0^y x^{a-1}\,(1-x)^{b-1}\,\mathrm{d}x]/[\int_0^1 x^{a-1}\,(1-x)^{b-1}\,\mathrm{d}x]$ is the regularized (in the sense that $I(1; a, b) = 1$) incomplete Beta function and $P_>(x) \equiv \int_x^\infty P(x')\,\mathrm{d}x'$ is the exceedance probability.

The main message from either the exact result in equation (2) or the extremal condition in equation (1) is that the time gap $g_k = t_{k+1} - t_k$ has the following generic behaviors (see appendix B for three simple examples):

- If $P(t)$ is constant, then $\langle g_k \rangle$ is independent of k.
- If $P(t)$ increases monotonically as t increases, then $\langle g_k \rangle$ decreases monotonically as k increases.
- If $P(t)$ decreases monotonically as t increases, then $\langle g_k \rangle$ increases monotonically as k increases.

Let us now apply the above results to marathon finishing times. As a trivial and artificial initial example, suppose that the marathon runners' speeds s are distributed

J. Stat. Mech. (2008) L03001

exponentially, $P(s) = s_*^{-1}\,\mathrm{e}^{-s/s_*}$, with s_* a characteristic running speed. Then the distribution of finishing times $t = L/s$ would be

$$P_0(t) = \frac{T}{t^2}\,\mathrm{e}^{-T/t}, \tag{3}$$

where $T = L/s_*$ is a typical finishing time for the field, and L is the course length. Applying the extremal criterion equation (1) to this distribution gives the typical kth finishing time $t_k = T/[\ln(N/k)]$. While t_k increases with k, as it must, this result has the unrealistic feature that the winning time approaches zero as the field becomes arbitrarily large.

More plausibly, the finishing time distribution should incorporate a non-zero fastest time $t_{\min}$. A slightly more refined example that obeys this constraint is

$$P(t) = \frac{mT^m}{\tau^{m+1}}\,\mathrm{e}^{-(T/\tau)^m}, \tag{4}$$

where $\tau = (t - t_{\min})$. The main new features of this distribution compared to equation (3) are the cutoff at $t_{\min}$ and the arbitrary exponent value m; the power-law prefactor is subdominant and it merely serves to simplify the calculations below. In fact, with the values $t_{\min} = 1.75$ h, $T = 2.75$ h and $m = 3$, equation (4) roughly follows the data in figure 1 (dashed curve). While one should not take the distribution (4) and the parameter values too seriously, we will see that its precise form does not affect the behavior of the time gaps between successive finishers.

Applying the extremal condition equation (1) to the distribution equation (4), and using the variable change $x = (T/\tau)^m$ to simplify the resulting integral, the typical value of the time gap is

$$g_k = T\left[\left(\ln\frac{N}{k+1}\right)^{-1/m} - \left(\ln\frac{N}{k}\right)^{-1/m}\right] \approx \frac{T}{m(\ln N)^{1+1/m}}\frac{1}{k} \qquad 1 \ll k \ll N. \tag{5}$$

This $1/k$ dependence holds for any distribution with an exponentially fast cutoff near the lower limit. The behavior $g_k \propto 1/k$ accords well with the data beyond approximately 20th place. However, contrary to the prediction from equation (5), the data clearly show that there is an 'excess' of elite runners (figure 2), as the time gaps between successive finishers are roughly constant for the first 20 places. Moreover, for the US races, the gaps between the first few consecutive finishers actually decrease with k. As seen in figure 2 for US races, the largest gap occurs between 5th and 6th place.

The reason that equation (5) does not capture the small-k behavior seen in figure 2 is that the parent distribution in equation (4) quickly goes to zero close to the fastest finishing time $t_{\min}$, whereas the actual distribution becomes nearly flat in this regime (figure 1). If we were to consider a flat distribution $P(t)$, as suggested by the data shown in figure 1, then a constant gap would be reproduced. The generic behavior of the dependence of the gap g_k on k is discussed in appendix B. Along these lines, a recent theory [4] predicts a crowding of runners near the front of marathon packs when the finishing time distribution is bounded from below. One additional feature of the gaps is that they begin to increase with $k \gtrsim 1000$ (figure 2). This behavior also arises from equation (5) for large k. This regime corresponds to finishing times of more than 4 h and is not relevant for our main conclusions.

J. Stat. Mech. (2008) L03001

Is there an explanation for having an excess of world-class runners? Many elite runners enjoy considerable incentives to maintain their competitive edge, including appearance money, access to the best support institutions (medical and athletic), etc. Thus if one achieves a time that qualifies as an elite performance, one is then in a position to take advantage of the various inducements offered to leading runners to maintain such a status. However, runners at the next tier of achievement face a daunting challenge. To run a marathon in the range, say, of 2:15–2:30 (for men) is still an impressive achievement that requires significant talent, dedication and time commitment. However, such a finish time is too slow to be competitive at major marathons. Thus runners who finish in this range typically have little or no external support for their athletic activities and have to balance this all-consuming endeavor with the need to survive economically. Consequently, one may even anticipate a deficit of male runners who can complete a marathon in the range of 2:15–2:30. Such a feature does actually occur in the Boston marathon.

It would be valuable to study whether a similar excess of the elite exists in different athletic events or other forms of human competition. It is also worth mentioning that perhaps a similar elite excess occurs in human mortality, where there is a well-known mortality plateau among the longest-lived individuals [5]–[7]. Here again, there seems to be a self-selected sub-population of advantaged individuals who gain advantage both innately and perhaps because of external reinforcement.

One of us (SR) thanks Guoan Hu for his invaluable data collection assistance, Paul Krapivsky for a helpful discussion, and the NSF for financial support. SS and SNM acknowledge the support of the Indo-French Center for the Promotion of Advanced Research under Project 3404-2.

Appendix A. Probability distribution of the kth finishing time

For a set of N iid random times that are drawn from the same distribution $P(t)$, let $\{t_1, t_2, \ldots, t_N\}$ denote their ordered set, with $t_1 < t_2 < \cdots < t_N$. Thus t_1 denotes the winning time, t_2 denotes the 2nd-place time, and t_k denotes the kth-fastest finishing time.

The probability distribution of the kth-fastest finishing time t_k is given by

$$\begin{aligned} f(t_k) &= \frac{N!}{(N-k)!(k-1)!}\left[\int_{t_k}^{\infty} P(x)\,\mathrm{d}x\right]^{N-k}\left[\int_0^{t_k} P(x)\,\mathrm{d}x\right]^{k-1} P(t_k), \\ &= \frac{1}{B(N-k+1,k)}\Big[P_>(t_k)\Big]^{N-k}\Big[1-P_>(t_k)\Big]^{k-1}\left[-\frac{\mathrm{d}P_>(t_k)}{\mathrm{d}t_k}\right]. \end{aligned} \quad \text{(A.1)}$$

Equation (A.1) merely specifies that $(N-k)$ variables are greater than t_k, $(k-1)$ variables are smaller than t_k, and one variable equals t_k. The combinatorial prefactor gives the number of such arrangements of these variables. In the second line, $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Beta function, and we have defined the exceedance probability

$$P_>(x) \equiv \int_x^{\infty} P(x')\,\mathrm{d}x',$$

namely, the probability that a variable chosen from the initial distribution P exceeds x. This exceedance probability satisfies the obvious conditions $P_>(0) = 1$ and $P_>(\infty) = 0$.

J. Stat. Mech. (2008) L03001

One can easily check from equation (A.1) that $f(t_k)$ is normalized, i.e., $\int_0^\infty f(t_k)\,\mathrm{d}t_k = 1$, as it must be.

The average value of the kth-fastest finishing time is then

$$\begin{aligned}\langle t_k\rangle &= \frac{1}{B(N-k+1,k)}\int_0^\infty x\left[P_>(x)\right]^{N-k}\left[1-P_>(x)\right]^{k-1}\left[-\frac{\mathrm{d}P_>(x)}{\mathrm{d}x}\right]\mathrm{d}x\\ &\equiv -\int_0^\infty x\,\frac{\mathrm{d}I}{\mathrm{d}x}\,\mathrm{d}x. \end{aligned}\tag{A.2}$$

In the second line we have introduced $I = I(y;a,b)$, the regularized incomplete Beta function, $I(y;a,b) \equiv B(y;a,b)/B(a,b)$, in which $B(y;a,b)$ is the incomplete Beta function

$$B(y;a,b) = \int_0^y x^{a-1}(1-x)^{b-1}\,\mathrm{d}x, \qquad y\in[0,1],$$

$B(a,b) = B(1;a,b)$ is the standard Beta function, and $y \equiv P_>(x)$. Integrating equation (A.2) by parts, and using the fact that the integrated term vanishes at both endpoints, gives the mean kth finishing time expressed by equation (2).

Appendix B. $\langle g_k\rangle$ for three simple cases

In this appendix we calculate $\langle g_k\rangle$ explicitly for three simple cases of $P(t)$.

Case 1. For the uniform distribution, $P(t) = 1$ in $t\in[0,1]$ and $P(t) = 0$ outside. Hence, $P_>(t) = 1-t$. Then

$$\begin{aligned}\langle t_k\rangle &= \frac{1}{B(N-k+1,k)}\int_0^1 x\cdot x^{k-1}(1-x)^{N-k}\,\mathrm{d}x = \frac{B(N-k+1,k+1)}{B(N-k+1,k)}\\ &= \frac{k}{N+1}. \end{aligned}\tag{B.1}$$

Thus we obtain $\langle g_k\rangle = 1/(N+1)$ for all k, while using the extremal condition (1) one finds the typical gap $g_k \approx 1/N$. As expected, $\langle g_k\rangle$ is independent of k for a uniform distribution.

Case 2. Consider the monotonically increasing distribution $p(t) = 2t$ in $t\in[0,1]$. Then $P_>(t) = 1-t^2$. Hence,

$$\begin{aligned}\langle t_k\rangle &= \frac{1}{B(N-k+1,k)}\int_0^1 x\cdot x^{2(k-1)}(1-x^2)^{N-k}\,\mathrm{d}x\\ &= \frac{1}{B(N-k+1,k)}\int_0^1 z^{k-1/2}(1-z)^{N-k}\,\mathrm{d}z\\ &= \frac{B(N-k+1,k+1/2)}{B(N-k+1,k)} = \frac{\Gamma(N+1)\Gamma(k+1/2)}{\Gamma(k)\Gamma(N+3/2)}. \end{aligned}\tag{B.2}$$

From this exact calculation we find

$$\begin{aligned}\langle g_k\rangle &= \frac{\Gamma(N+1)}{\Gamma(N+3/2)}\frac{\Gamma(k+1/2)}{\Gamma(k)}\frac{1}{2k}, \qquad \text{for all } N \qquad \text{and}\\ &\approx \frac{1}{2\sqrt{N}}\frac{1}{\sqrt{k}}, \qquad \text{for } N > k \gg 1. \end{aligned}\tag{B.3}$$

J. Stat. Mech. (2008) L03001

Similarly using equation (1) the typical value of the kth finishing time is $t_k \approx \sqrt{k/N}$, and hence

$$g_k \approx \frac{1}{\sqrt{N}}\left[\sqrt{k+1}-\sqrt{k}\right] \approx \frac{1}{2\sqrt{N}}\frac{1}{\sqrt{k}}, \qquad \text{for } N > k \gg 1. \tag{B.4}$$

These results show that $\langle g_k \rangle$ monotonically decreases as k increases. That is, the gap between successive variables gets smaller when their density increases, as one would expect.

Case 3. Consider the monotonically decreasing distribution $P(t) = \exp(-t)$ where $t \in [0, \infty)$. In this case,

$$\langle t_k \rangle = \frac{1}{B(N-k+1,k)}\int_0^\infty x\mathrm{e}^{-(N-k)x}(1-\mathrm{e}^{-x})^{k-1}\,\mathrm{d}x = -\frac{1}{B(N-k+1,k)}\int_0^1 \ln z^{N-k}(1-z)^{k-1}\,\mathrm{d}z.$$

The latter integral can be found in [8] and the final result is

$$\langle t_k \rangle = \psi(N+1) - \psi(N-k+1), \tag{B.5}$$

where $\psi(x) = \mathrm{d}\ln\Gamma(x)/\mathrm{d}x$ is the digamma function. Finally using the series representation

$$\psi(n) = -\gamma + \sum_{m=1}^{n-1}\frac{1}{m}, \tag{B.6}$$

where $\gamma = 0.577\,215\ldots$ is Euler's constant, we obtain

$$\langle g_k \rangle = \frac{1}{N-k}, \qquad 1 \le k \le N-1. \tag{B.7}$$

On the other hand using the extremal condition (equation (1)) one finds the typical value

$$g_k \approx -\log\left(1-\frac{1}{N-k}\right) \approx \frac{1}{N-k}, \qquad \text{for } N-k \gg 1. \tag{B.8}$$

Thus $\langle g_k \rangle$ monotonically increases as k increases. Note that in this case the extremal condition equation (1) does not describe well the behavior when k is close to N.

References

[1] The data reported here can be obtained from http://www.marathonguide.com
[2] Fisher R A and Tippett L H C, *Limiting forms of the frequency distribution of the largest or smallest member of a sample*, 1928 *Proc. Camb. Phil. Soc.* **24** 180
[3] Gumbel E J, 1958 *Statistics of Extremes* (New York: Columbia University Press)
[4] Sabhapandit S and Majumdar S N, *Density of near-extreme events*, 2007 *Phys. Rev. Lett.* **98** 140201
[5] Greenwood M and Irwin J O, *The biostatistics of senility*, 1939 *Hum. Biol.* **11** 1
[6] Penna T J P, *A bit-string model for biological aging*, 1995 *J. Stat. Phys.* **78** 1629
[7] Azbel M Ya, *Phenomenological theory of mortality evolution: its singularities, universality, and superuniversality*, 1999 *Proc. Nat. Acad. Sci.* **96** 3303
[8] Gradshteyn I S and Ryzhik I M, 2007 *Table of Integrals, Series and Products* 7th edn, ed A Jeffrey and D Zwillinger (New York: Academic)

J. Stat. Mech. (2008) L03001