The Annals of Applied Probability
2014, Vol. 24, No. 4, 1375–1395
DOI: 10.1214/13-AAP949

ON THE EXISTENCE OF ACCESSIBLE PATHS IN VARIOUS MODELS OF FITNESS LANDSCAPES

BY PETER HEGARTY AND ANDERS MARTINSSON

Chalmers University of Technology and University of Gothenburg

We present rigorous mathematical analyses of a number of well-known mathematical models for genetic mutations. In these models, the genome is represented by a vertex of the n-dimensional binary hypercube, for some n, a mutation involves the flipping of a single bit, and each vertex is assigned a real number, called its fitness, according to some rules. Our main concern is with the issue of existence of (selectively) accessible paths; that is, monotonic paths in the hypercube along which fitness is always increasing. Our main results resolve open questions about three such models, which in the biophysics literature are known as house of cards (HoC), constrained house of cards (CHoC) and rough Mount Fuji (RMF). We prove that the probability of there being at least one accessible path from the all-zeroes node $\mathbf{v}^0$ to the all-ones node $\mathbf{v}^1$ tends respectively to 0, 1 and 1, as n tends to infinity. A crucial idea is the introduction of a generalization of the CHoC model, in which the fitness of $\mathbf{v}^0$ is set to some $\alpha = \alpha_n \in [0, 1]$. We prove that there is a very sharp threshold at $\alpha_n = \frac{\ln n}{n}$ for the existence of accessible paths from $\mathbf{v}^0$ to $\mathbf{v}^1$. As a corollary we prove significant concentration, for α below the threshold, of the number of accessible paths about the expected value (the precise statement is technical; see Corollary 1.4). In the case of RMF, we prove that the probability of accessible paths from $\mathbf{v}^0$ to $\mathbf{v}^1$ existing tends to 1 provided the drift parameter $\theta = \theta_n$ satisfies $n\theta_n \to \infty$, and for any fitness distribution which is continuous on its support and whose support is connected.

0. Notation. Throughout this paper, $\mathbb{Q}_n$ will denote the *directed* n-dimensional binary hypercube. This is the directed graph whose nodes are all binary strings of length n, with an edge between any pair of nodes that differ in exactly one bit, the edge being always directed toward the node with the greater number of ones.

Let $g, h : \mathbb{N} \to \mathbb{R}_+$ be any two functions. We will employ the following notation throughout, all of which is quite standard:

(i) $g(n) \sim h(n)$ means that $\lim_{n\to\infty} \frac{g(n)}{h(n)} = 1$;
(ii) $g(n) \lesssim h(n)$ means that $\limsup_{n\to\infty} \frac{g(n)}{h(n)} \leq 1$;
(iii) $g(n) \gtrsim h(n)$ means that $h(n) \lesssim g(n)$;

Received October 2012; revised July 2013.
MSC2010 subject classifications. Primary 60C05, 92D15; secondary 05A05.
Key words and phrases. Accessible path, hypercube, percolation, house of cards, rough Mount Fuji.

(iv) $g(n) = O(h(n))$ means that $\limsup_{n\to\infty} \frac{g(n)}{h(n)} < \infty$;
(v) $g(n) = \Omega(h(n))$ means that $h(n) = O(g(n))$;
(vi) $g(n) = \Theta(h(n))$ means that both $g(n) = O(h(n))$ and $h(n) = O(g(n))$ hold;
(vii) $g(n) = o(h(n))$ means that $\lim_{n\to\infty} \frac{g(n)}{h(n)} = 0$.

Now suppose instead that $(g(n))_{n=1}^{\infty}, (h(n))_{n=1}^{\infty}$ are two sequences of random variables. We write $g(n) \sim h(n)$ if, for all $\varepsilon_1, \varepsilon_2 > 0$ and n sufficiently large,

$$\mathbb{P}\left(1 - \varepsilon_1 < \frac{g(n)}{h(n)} < 1 + \varepsilon_1\right) > 1 - \varepsilon_2. \tag{0.1}$$

Similarly, we write $g(n) \gtrsim h(n)$ if, for all $\varepsilon_1, \varepsilon_2 > 0$ and n sufficiently large,

$$\mathbb{P}\left(\frac{g(n)}{h(n)} > 1 - \varepsilon_1\right) > 1 - \varepsilon_2. \tag{0.2}$$

1. Introduction. In many basic mathematical models of genetic mutations, the genome is represented as a node of the directed n-dimensional binary hypercube $\mathbb{Q}_n$, and each mutation involves the flipping of a single bit from 0 (the "wild" state) to 1 (the "mutant" state), hence displacement along an edge of $\mathbb{Q}_n$. Each node $v \in \mathbb{Q}_n$ is assigned a real number $f(v)$, called its *fitness*. The fitness of a node is not a constant, but is drawn from some probability distribution specified by the model. This distribution may vary from node to node in more or less complicated ways, depending on the model. Basically, however, evolution is considered as favoring mutational pathways which, on average, lead to higher fitness. A fundamental concept in this regard is the following (see [6, 15, 16]):

DEFINITION 1.1. Let $f : \mathbb{Q}_n \to \mathbb{R}$ be a fitness function. A *(selectively) accessible path* in $\mathbb{Q}_n$ is a path

$$v_0 \to v_1 \to \cdots \to v_{k-1} \to v_k, \tag{1.1}$$

such that $f(v_i) > f(v_{i-1})$ for $i = 1, \ldots, k$.

Let $\mathbf{v}^0 = (0, 0, \ldots, 0)$, $\mathbf{v}^1 = (1, 1, \ldots, 1)$ denote the all-zeroes and all-ones vertices in $\mathbb{Q}_n$. A basic question in such models is whether accessible paths from $\mathbf{v}^0$ to $\mathbf{v}^1$ exist or not with high probability. For the remainder of this paper, unless explicitly stated otherwise, the words "accessible path" will always refer to such a path which starts at $\mathbf{v}^0$ and ends at $\mathbf{v}^1$. In fact, it will only be in the proof of Proposition 2.18 that we will need to consider accessible paths with other start- and endpoints.

We shall be concerned below with the following three well-known models, in which no rigorous answer has previously been given to the question of whether or not accessible paths exist with high probability.

MODEL 1 [Unconstrained house of cards (HoC)]. This model is originally attributed to Kingman [10]. In the form we consider below, it was first studied by Kauffman and Levin [9]. We set $f(\mathbf{v}^1) := 1$ and, for every other node $v \in \mathbb{Q}_n$, independently let $f(v) \sim \mathrm{U}(0, 1)$, the uniform distribution on the interval $[0, 1]$.

MODEL 2 [Constrained house of cards (CHoC)]. This variant seems to have been considered only more recently; see, for example, [11] and [3]. The only difference from Model 1 is that we fix $f(\mathbf{v}^0) := 0$.

MODEL 3 [Rough Mount Fuji (RMF)]. This model was first proposed in [1]; see also [8]. For each $v \in \mathbb{Q}_n$, one lets

$$f(v) = \theta \cdot d\big(v, \mathbf{v}^0\big) + \eta(v), \tag{1.2}$$

where $\theta = \theta_n$ is a positive number called the *drift*, $d(\cdot, \cdot)$ denotes Hamming distance and the $\eta(v)$ are independent random variables of some fixed distribution. In other words, one first assigns a fitness to each node at random, according to η and independent of all other nodes. Then the fitness of each node is shifted upward by a fixed multiple of the Hamming distance from $\mathbf{v}^0$.

Before proceeding, it is worth noting that the above models are also of interest in physics in the context of so-called *spin glasses* [12]. In this setting, each node of $\mathbb{Q}_n$ represents a point in the state space of all possible configurations of spins in a disordered magnet. The analogue of fitness is in this case energy, or more precisely "energy times -1." Accessible paths (not necessarily from $\mathbf{v}^0$ to $\mathbf{v}^1$) correspond to trajectories in which energy decreases monotonically, and which are therefore easily accessible even at zero temperature. The HoC model appears in the spin glass context as Derrida's random energy model (REM), and the RMF-model is a REM in an external magnetic field. For further discussion of the connection between fitness landscapes and spin glasses, see [7].

In all three models, the basic random variable of interest is the number $X = X(n)$ of accessible paths. One thinks of $\mathbf{v}^0$ as the starting point of some evolutionary process, and $\mathbf{v}^1$ as the desirable endpoint. The HoC model is often referred to as a "null model" for evolution, since the fitnesses of all nodes other than $\mathbf{v}^1$ are assigned at random and independently of one another. No mechanism is prescribed which might push an evolutionary process in any particular direction. The CHoC model is not much better, though it does specify that the starting point is a global fitness minimum. The RMF model is a very natural and simple way to introduce an "arrow of evolution," since the drift factor implies that successive $0 \to 1$ mutations will tend to increase fitness.

It seems intuitively obvious that the number X of accessible paths should, on average, be much higher in RMF than in HoC, with the CHoC model lying somewhere in between. One should be a little careful here, since in RMF, the node $\mathbf{v}^1$ is not assumed to be a global fitness maximum. Nevertheless, it is easy to verify that

$\mathbb{E}[X] = 1$ in HoC, $\mathbb{E}[X] = n$ in CHoC, whereas in many situations $\mathbb{E}[X]$ grows super-exponentially with n in RMF; see [6], along with Propositions 2.1 and 3.1 below. Of more interest, however, is the quantity $P = P(n)$, which is the probability of there being at least one accessible path, that is, $P = \mathbb{P}(X > 0)$. The idea here is that, as long as *some* accessible path exists, then evolution will eventually find it. The quantity P has been simulated in the biophysics literature. In [6] it was conjectured explicitly that $P \to 0$ in the HoC model, and that $P \to 1$ in the RMF model, when η is a normal distribution and θ is any positive constant. In [3], the CHoC model was simulated for $n \leq 13$, and the authors conjecture, if somewhat implicitly, that P is monotonic decreasing in n and approaches a limiting value close to 0.7. In [6], simulations were continued up to $n = 19$, and these indicated clearly that P was not, after all, monotonic decreasing. The authors abstain from making any explicit conjecture about the limiting behavior of P in CHoC.

Our main results below resolve all these issues. A crucial idea is to consider the following slight generalization of the CHoC model:

MODEL 4 [α-Constrained House of Cards (α-HoC)]. Let $\alpha \in [0, 1]$. In this model, fitnesses are assigned as in the CHoC model, with the exception that we set $f(\mathbf{v}^0) := \alpha$. Hence, CHoC is the case $\alpha = 0$.

For $\alpha \in [0, 1]$, let $P(n, \alpha)$ denote the probability of there being an accessible path in the α-HoC model. To simplify notation below, we define $P(n, \alpha) = P(n, 0)$ for $\alpha < 0$ and $P(n, \alpha) = P(n, 1)$ for $\alpha > 1$. Note that $P(n, \alpha)$ decreases as α increases. Our first main result is the following:

THEOREM 1.2. *Let $\varepsilon = \varepsilon_n > 0$. If $n\varepsilon_n \to \infty$, then*

$$\lim_{n\to\infty} P\left(n, \frac{\ln n}{n} - \varepsilon_n\right) = 1 \tag{1.3}$$

and

$$\lim_{n\to\infty} P\left(n, \frac{\ln n}{n} + \varepsilon_n\right) = 0. \tag{1.4}$$

It follows immediately that $P \to 1$ in the CHoC model and that $P(n, \alpha) \to 0$ for any strictly positive constant α. The above result says a lot more, however. It shows that there is a very sharp threshold at $\alpha = \alpha_n = \frac{\ln n}{n}$ for the existence of accessible paths in the α-HoC model. Theorem 1.2 will be proven in Section 2. We have the following immediate corollary for HoC:

COROLLARY 1.3. *Let X denote the number of accessible paths in the HoC model. Then*

$$\mathbb{P}(X > 0) \sim \frac{\ln n}{n}. \tag{1.5}$$

PROOF. As $P(n,\alpha)$ is decreasing in α we know that, for any $\alpha \in [0,1]$, $\mathbb{P}(X>0) \geq \alpha P(n,\alpha)$. Picking $\alpha = \frac{\ln n}{n} - \varepsilon_n$ where $n\varepsilon_n$ tends to infinity sufficiently slowly, it follows from Theorem 1.2 that $\mathbb{P}(X>0) \gtrsim \frac{\ln n}{n}$.

To get the upper bound, let $\alpha = \frac{\ln n}{n}$. Now if the hypercube has accessible paths, then either $\mathbf{v}^0$ has fitness at most α, or there is an accessible path where all nodes involved have fitness at least α. Obviously the former event occurs with probability α. Concerning the latter, if

$$\mathbf{v}^0 \to v_1 \to \cdots \to v_{n-1} \to \mathbf{v}^1 \tag{1.6}$$

is any path, then the probability of all nodes along it having fitness at least α is $(1-\alpha)^n$. The probability of fitness being increasing along the path is $1/n!$. Since there are $n!$ possible paths of the form (1.6), it follows from a union bound that

$$\mathbb{P}(X>0) \leq \alpha + n!\frac{(1-\alpha)^n}{n!} \leq \frac{\ln n}{n} + \frac{1}{n}. \tag{1.7}$$

□

Another corollary of Theorem 1.2 concerns the distribution of the number of accessible paths in α-HoC for $\alpha = \frac{\ln n}{n} - \varepsilon_n$, where $n\varepsilon_n \to \infty$. It is straightforward to show that the expected number of paths in α-HoC is $n(1-\alpha)^{n-1}$ (see Proposition 2.1), which, for this choice of α, is $\sim e^{n\varepsilon_n}$. We have the following result:

COROLLARY 1.4. *Let X denote the number of accessible paths in α-HoC for $\alpha = \frac{\ln n}{n} - \varepsilon_n$ where $n\varepsilon_n \to \infty$. If $w_n \to \infty$, then*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{w_n}\mathbb{E}[X] \leq X \leq w_n \mathbb{E}[X]\right) = 1. \tag{1.8}$$

Corollary 1.4 will be proven in Section 2.5.

Our second main result concerns the RMF model. For any function $f:\mathbb{R}\to\mathbb{R}$, recall that the *support* of f, denoted $\mathrm{Supp}(f)$, is the set of points at which f is nonzero,[1] that is, $\mathrm{Supp}(f) = \{x : f(x) \neq 0\}$. We say that f has *connected support* if $\mathrm{Supp}(f)$ is a connected subset of $\mathbb{R}$. Our result is the following:

THEOREM 1.5. *Let η be any probability distribution whose p.d.f. is continuous on its support and whose support is connected. Let θ_n be any strictly positive function of n such that $n\theta_n \to \infty$ as $n \to \infty$. Then in the model (1.2), $P(n)$ tends to one as $n \to \infty$.*

This result is proven in Section 3. The proof follows similar lines to that of Theorem 1.2, but the analysis is somewhat simpler.

[1] Sometimes in the mathematical literature, the support of a function is defined to be the closure of this set.

REMARK 1.6. More generally, the proof of Theorem 1.5 presented in this article holds for any distribution η that satisfies, with notation taken from Section 3, $\kappa_{\eta,\delta} = \inf_{I \subseteq I_\delta} \frac{1}{l(I)} \int_I \eta(x)\,dx > 0$ for any $\delta \in (0,1)$. This condition essentially states that η is not allowed to have "isolated modes." For instance, it is satisfied for any unimodal distribution.

2. Results for the HoC models. For each path i from $\mathbf{v}^0$ to $\mathbf{v}^1$ let X_i be the indicator function of the event that i is accessible, and let $X = \sum_i X_i$ denote the number of accessible paths from $\mathbf{v}^0$ to $\mathbf{v}^1$. Furthermore, given a path i from $\mathbf{v}^0$ to $\mathbf{v}^1$ in the n-dimensional hypercube, let $T(n,k)$ denote the number of paths from $\mathbf{v}^0$ to $\mathbf{v}^1$ that intersect i in exactly $k-1$ interior nodes (by symmetry, this is independent of i).

PROPOSITION 2.1. *Let X denote the number of accessible paths in the α-HoC model. Then*

$$\mathbb{E}[X] = n(1-\alpha)^{n-1}. \tag{2.1}$$

PROOF. There are $n!$ paths through the hypercube. A path is accessible if all $n-1$ interior nodes have fitness at least α, and the fitness of the interior nodes is increasing along the path. This occurs with probability $(1-\alpha)^{n-1}/(n-1)!$. □

Note that for $\alpha = \frac{\ln n}{n} + \varepsilon_n$, the proposition implies that the expected number of accessible paths tends to 0 for any sequence ε_n satisfying $n\varepsilon_n \to \infty$. This directly implies equation (1.4). Similarly, for $\alpha = \frac{\ln n}{n} - \varepsilon_n$ where $n\varepsilon_n \to \infty$, the expected number of paths tends to infinity.

To show the remaining part of Theorem 1.2, that the probability of there being at least one accessible path tends to 1 in the case $\alpha = \frac{\ln n}{n} - \varepsilon_n$, we will begin by showing that the probability is at least $\frac{1}{4} - o(1)$ by the second moment method. In Section 2.4 we will then provide a proof that the probability must tend to 1.

LEMMA 2.2. *Let X be a random variable with finite expected value and finite and nonzero second moment. Then*

$$\mathbb{P}(X \neq 0) \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}. \tag{2.2}$$

PROOF. Let $1_{X\neq 0}$ denote the indicator function of $X \neq 0$. Then, by the Cauchy–Schwarz inequality, $\mathbb{E}[X]^2 = \mathbb{E}[1_{X\neq 0}X]^2 \leq \mathbb{E}[1_{X\neq 0}^2] \cdot \mathbb{E}[X^2] = \mathbb{P}(X \neq 0) \cdot \mathbb{E}[X^2]$. □

See also Exercise 4.8.1 in [2].

PROPOSITION 2.3. *Let i and j be paths with exactly $k-1$ interior nodes in common. Then*

$$\mathbb{E}[X_i X_j] \le \frac{\binom{2n-2k}{n-k}(1-\alpha)^{2n-k-1}}{(2n-k-1)!}, \tag{2.3}$$

where equality holds if the nodes where i and j differ are consecutive along the paths, that is, if i and j diverge at most once. Furthermore,

$$\mathbb{E}[X^2] \le \sum_{k=1}^{n} n!T(n,k)\frac{\binom{2n-2k}{n-k}(1-\alpha)^{2n-k-1}}{(2n-k-1)!}. \tag{2.4}$$

PROOF. The event that i and j are both accessible occurs if all $2n-k-1$ interior nodes have fitness at least α and the fitnesses of the interior nodes are ordered in such a way that fitness increases along both paths.

Conditioned on the event that all interior nodes have fitness at least α, all possible ways in which the fitnesses of the interior nodes can be ordered are equally likely. This implies that the probability that both paths are accessible is $(1-\alpha)^{2n-k-1}/(2n-k-1)!$ times the number of ways to order the fitnesses of the interior nodes such that fitness increases along both paths.

To count the number of ways this can be done we color the numbers $1,\dots,2n-k-1$ in the following way: The number l is colored gray if the interior node with the lth smallest fitness is contained in both paths, red if it is only contained in i and blue if only in j. Note that i and j uniquely determine which numbers must be gray for a valid order, and that any coloring corresponds to at most one order.

Clearly, any coloring corresponding to a valid order colors half of the nongray numbers red and half blue, which implies that there can be at most $\binom{2n-2k}{n-k}$ such orders. Furthermore, if i and j diverge at most once, one can always construct a valid order from such a coloring, so in this case there are exactly $\binom{2n-2k}{n-k}$ such orders.

As the number of ordered pairs of paths that intersect in exactly $k-1$ interior nodes is $n!T(n,k)$, (2.4) follows from this estimate. □

2.1. *Useful formulas for $T(n,k)$.* The numbers $T(n,k)$ already appear in the mathematical literature. The usual terminology is that $T(n,k)$ is the number of permutations of $\{1,2,\dots,n\}$ with k *components*, where the number of components of a permutation $\pi_1\pi_2\cdots\pi_n$ is defined as the number of choices for $1 \le s \le n$ such that $\pi_1\pi_2\cdots\pi_s$ is a permutation of $\{1,2,\dots,s\}$. In terms of paths in $\mathbb{Q}_n$, we can represent each path from $\mathbf{v}^0$ to $\mathbf{v}^1$ by a permutation $\pi_1\pi_2\cdots\pi_n$ of $\{1,2,\dots,n\}$ where π_s denotes which coordinate to increase in step s. If we let i be the path represented by the identity permutation, then a path j, represented by $\pi_1\pi_2\cdots\pi_n$, intersects i in step $s \ge 1$ if and only if $\pi_1\pi_2\cdots\pi_s$ is a permutation of $\{1,2,\dots,s\}$.

This means that, if $\pi_1\pi_2\cdots\pi_n$ has k components, then i and j intersect in $k-1$ interior nodes (the kth component corresponds to $s=n$). We can thus consider a component as an interval $[s,t]$ where i and j intersect in steps s and t, but at no step in between.

An alternative formulation is that $T(n,k)$ is the number of permutations of $\{1,2,\dots,n\}$ with $k-1$ *global descents*. A global descent in a permutation $\pi_1\pi_2\cdots\pi_n$ of $\{1,2,\dots,n\}$ is a number $t\in[1,n-1]$ such that $\pi_i>\pi_j$ for all $i\le t$ and $j>t$. There is a simple 1–1 correspondence between permutations with k components and those with $k-1$ global descents obtained by reading a permutation backward. In other words, $\pi_1\pi_2\cdots\pi_n$ has $k-1$ global descents if and only if $\pi_n\pi_{n-1}\cdots\pi_1$ has k components.

There is a database of the numbers $T(n,k)$ for small n and k; see [14]. Comtet's book [5] contains a couple of exercises and an implicit recursion formula for $T(n,k)$. Comtet has also performed a detailed asymptotic analysis of the numbers $T(n,1)$ in [4]. Permutations with one component (i.e., no global descents) are variously referred to as *connected, indecomposable, irreducible*. These seem to crop up quite a lot; see [13]. However, estimates of the numbers $T(n,k)$ for general n and k like those in Propositions 2.9 and 2.11 below do not appear to have been obtained before.

PROPOSITION 2.4. *$T(n,1)$ is uniquely defined by*

$$n!=\sum_{k=1}^{n}T(k,1)(n-k)!. \tag{2.5}$$

PROOF. Given a path i through $\mathbb{Q}_n$, the number of paths j that intersect i for the first time in step k is $T(k,1)(n-k)!$. As any path through $\mathbb{Q}_n$ intersects i for the first time after between 1 and n steps, the proposition follows. □

PROPOSITION 2.5.

$$n!\left(1-O\left(\frac{1}{n}\right)\right)\le T(n,1)\le n!. \tag{2.6}$$

PROOF. By definition, $T(n,1)\le n!$. Using this, Proposition 2.4 implies that $T(n,1)$ is at least $n!-\sum_{k=1}^{n-1}k!(n-k)!=n!-O((n-1)!)$. □

PROPOSITION 2.6.

$$T(n,k)=\sum_{\substack{s_1,\dots,s_k\ge1\\ s_1+\cdots+s_k=n}}T(s_1,1)\cdots T(s_k,1). \tag{2.7}$$

PROOF. Given a path i, the number of paths that intersect i for the first time after s_1 steps, for the second time after s_2 more steps and so on up to the last time (at $\mathbf{v}^1$) after n steps is $T(s_1,1)\cdots T(s_{k-1},1)\cdot T(n-s_1-\cdots-s_{k-1},1)$. Let $s_k = n-s_1-\cdots-s_{k-1}$. $T(n,k)$ is obtained by summing over all possible values of $s_1,\ldots,s_k$. □

PROPOSITION 2.7. *For $k\geq 2$, $T(n,k)$ satisfies*

$$T(n,k)=\sum_{s=1}^{n-k+1} T(s,1)T(n-s,k-1). \tag{2.8}$$

PROOF. It follows by induction that this sum equals the right-hand side in (2.7). □

2.2. *Upper bounds for $T(n,k)$.*

PROPOSITION 2.8. *For any $n\geq k\geq 1$,*

$$T(n,k)\leq k\sum\left(\left(n-\sum_{j=1}^{k-1}s_j\right)!\prod_{j=1}^{k-1}s_j!\right), \tag{2.9}$$

where the first sum goes over all $(k-1)$-tuples of integers $s_1,\ldots,s_{k-1}$ such that $s_j\geq 1$ for all j and $\max_j s_j\leq n-\sum_j s_j$.

PROOF. Consider the formula for $T(n,k)$ in Proposition 2.6. By symmetry, $T(n,k)$ is at most k times the contribution from terms where $s_j\leq s_k$ for $j=1,\ldots,k-1$. The proposition follows by applying $T(s,1)\leq s!$. □

PROPOSITION 2.9. *There is a positive constant c such that for all $n\geq k\geq 1$,*

$$T(n,k)\leq k(n-k+1)!e^{c(k-1)/(n-k+1)}. \tag{2.10}$$

PROOF. We use Proposition 2.8 and make the following approximations:

- Substitute $(n-\sum_j s_j)!$ by $\beta^{n-\sum_j s_j}$ where $\beta=((n-k+1)!)^{1/(n-k+1)}$. It follows from log-convexity of $l!$ that $\beta^l\geq l!$ for any $0\leq l\leq n-k+1$.
- Let all s_j go from 1 to $\lfloor (n-k+1)/2+1\rfloor$.

This yields

$$T(n,k)\leq k(n-k+1)!\left(\sum_{s=1}^{\lfloor (n-k+1)/2+1\rfloor} s!\beta^{1-s}\right)^{k-1}. \tag{2.11}$$

We now claim that the sum in the above expression is always less than $1 + c/(n-k+1)$ for sufficiently large c. Indeed,

$$\begin{aligned}
&\sum_{s=1}^{\lfloor (n-k+1)/2+1 \rfloor} s!\beta^{1-s} \\
&= 1 + 2\beta^{-1} + \beta^{-1} \sum_{t=1}^{\lfloor (n-k+1)/2-1 \rfloor} t!(t+1)(t+2)\beta^{-t} \\
&\le 1 + 2\beta^{-1} \\
&\quad + e\beta^{-1} \sum_{t=1}^{\lfloor (n-k+1)/2-1 \rfloor} \sqrt{t}(t+1)(t+2)\left(\frac{n-k+1}{2e}\right)^{t}\left(\frac{n-k+1}{e}\right)^{-t} \\
&\le 1 + 2\beta^{-1} + e\beta^{-1} \sum_{t=1}^{\infty} \sqrt{t}(t+1)(t+2)2^{-t} \\
&\le 1 + c(n-k+1)^{-1}.
\end{aligned}$$

Here we have used that $(n-k+1)/e \le \beta \le (n-k+1)$ and that $n! \le en^{n+1/2}e^{-n}$, which follows from standard estimates of factorials.

The proposition now follows from this result together with (2.11). □

PROPOSITION 2.10. *For any fixed l there is a constant $C_l > 0$ such that*

$$T(n, n-l) \le C_l n^l \tag{2.12}$$

for all $n \ge 1$.

PROOF. We may, without loss of generality, assume that $n \ge 2l$.

Recall the formula for $T(n, n-l)$ in Proposition 2.6. As $s_1, \dots, s_{n-l} \ge 1$ and $s_1 + \cdots + s_{n-l} = n$ it is easy to see that all but at most l variables are equal to 1. This implies that $T(n, n-l)$ is at most $\binom{n-l}{l}$ times the contribution from all terms where $s_{l+1} = \cdots = s_{n-l} = 1$. Using $T(1,1) = 1$, we get

$$T(n, n-l) \le \binom{n-l}{l} \sum_{\substack{s_1,\dots,s_l \ge 1 \\ s_1+\cdots+s_l = 2l}} T(s_1, 1) \cdots T(s_l, 1) \le C_l n^l. \tag{2.13}$$

□

PROPOSITION 2.11. *For sufficiently large c, we have*

$$T(n, n-l) \le c(l+1)\left(\frac{n+2l}{5}\right)^{l}. \tag{2.14}$$

PROOF. Let

$$S(n, n-l) = (l+1)\left(\frac{n+2l}{5}\right)^l, \tag{2.15}$$

that is,

$$S(n,k) = (n-k+1)\left(\frac{3n-2k}{5}\right)^{n-k}. \tag{2.16}$$

We will begin by showing that $S(n,k)$ satisfies

$$S(n,k) \geq \sum_{i=1}^{n-k+1} i!S(n-i,k-1) \tag{2.17}$$

for $k > 1$ and sufficiently large $n-k$. Here we have

$$\begin{aligned}
&\sum_{i=1}^{n-k+1} i!S(n-i,k-1) \\
&\quad = \sum_{i=1}^{n-k+1} i!(n-k+2-i)\left(\frac{3n-2k-3i+2}{5}\right)^{n-k-i+1} \\
&\quad \leq (n-k+1)\left(\frac{3n-2k-1}{5}\right)^{n-k} \\
&\qquad + \sum_{i=2}^{n-k+1} i!(n-k+1)\left(\frac{3n-2k}{5}\right)^{n-k-i+1} \\
&\quad = S(n,k)\left(\left(1-\frac{1}{3n-2k}\right)^{n-k} + \sum_{i=2}^{n-k+1} i!\left(\frac{3n-2k}{5}\right)^{-i+1}\right),
\end{aligned}$$

where

$$\begin{aligned}
\left(1-\frac{1}{3n-2k}\right)^{n-k} &\leq \exp\left(-\frac{n-k}{3n-2k}\right) \\
&\leq \exp\left(-\frac{n-k}{3n}\right) \leq \max\left(\frac{1}{2}, 1-\frac{n-k}{6n}\right)
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{i=2}^{n-k+1} i!\left(\frac{3n-2k}{5}\right)^{-i+1} \\
&\quad \leq \frac{10}{3n-2k} + \frac{5}{3n-2k}\sum_{j=1}^{n-k-1} j!(j+1)(j+2)\left(\frac{3n-2k}{5}\right)^{-j}
\end{aligned}$$

$$\leq \frac{10}{3n-2k} + \frac{5e}{3n-2k}\sum_{j=1}^{\infty}\sqrt{j}(j+1)(j+2)\left(\frac{n-k}{e}\right)^{j}\left(\frac{3n-2k}{5}\right)^{-j}$$

$$\leq \frac{1}{n}\left(10 + 5e\sum_{j=1}^{\infty}\sqrt{j}(j+1)(j+2)\left(\frac{5}{3e}\right)^{j}\right)$$

$$= \frac{C}{n}.$$

It follows directly that (2.17) holds for $k > 1$ and $n - k \geq 6C$.

Now, if we can choose c so that $T(n,k) \leq cS(n,k)$ for $k = 1$ and for $n - k < 6C$, the proposition will follow from Proposition 2.7 by induction on k. Hence it suffices to show the proposition for these two cases.

For $k = 1$, the inequality holds for sufficiently large c by the fact that

$$\begin{aligned}\frac{T(n,1)}{S(n,1)} &\leq \frac{n!}{n((3n-2)/5)^{n-1}}\\ &\leq e\sqrt{n}\left(\frac{n}{e}\right)^{n}\frac{1}{n((3n-2)/5)^{n-1}}\\ &= \frac{3e}{5}\sqrt{n}\left(\frac{5}{3e}\right)^{n}\left(1-\frac{2}{3n}\right)^{-n+1}\\ &\to 0 \qquad \text{as } n\to\infty.\end{aligned}$$

For $n - k < 6C$, just apply Proposition 2.10. □

2.3. *Computing* $\mathbb{E}[X^2]$. Pick $\delta > 0$ sufficiently small. We divide the sum in (2.4) into the contribution from $k \leq (1-\delta)n$ and that from $k > (1-\delta)n$:

$$\begin{aligned}&\sum_{k=1}^{n} n!T(n,k)\frac{\binom{2n-2k}{n-k}(1-\alpha)^{2n-k-1}}{(2n-k-1)!}\\ &\quad= \sum_{k=1}^{(1-\delta)n} n!T(n,k)\frac{\binom{2n-2k}{n-k}(1-\alpha)^{2n-k-1}}{(2n-k-1)!}\\ &\qquad+ \sum_{l=0}^{\delta n} n!T(n,n-l)\frac{\binom{2l}{l}(1-\alpha)^{n+l-1}}{(n+l-1)!}\\ &\quad:= S_1 + S_2.\end{aligned} \tag{2.18}$$

PROPOSITION 2.12. *For k constant and $\alpha = o(1)$*

$$n!T(n,k)\frac{\binom{2n-2k}{n-k}(1-\alpha)^{2n-k-1}}{(2n-k-1)!} \sim k2^{1-k}n^2(1-\alpha)^{2n}. \tag{2.19}$$

PROOF. A simple lower bound on $T(n,k)$ is the number of permutations with k components where all but one component contains exactly one element. For sufficiently large n this is given by $kT(n-k+1,1)$, which by Proposition 2.5 is $\sim k(n-k+1)!$. Furthermore, from Proposition 2.9 we know that $T(n,k)$ is most $(1+o(1))k(n-k+1)!$. Hence for constant k, $T(n,k)\sim k(n-k+1)!$. The proposition now follows from standard estimates of factorials. □

PROPOSITION 2.13. *Let $\alpha = o(1)$. For any $0<\delta<1$, we have $S_1 \sim 4n^2(1-\alpha)^{2n}$.*

PROOF. From Proposition 2.9 it follows that there is a constant C_δ such that $T(n,k)\le C_\delta k(n-k+1)!$ whenever $k\le(1-\delta)n$. Using this we have

$$n!T(n,k)\frac{\binom{2n-2k}{n-k}}{(2n-k-1)!} \le C_\delta n!k(n-k+1)!\frac{\binom{2n-2k}{n-k}}{(2n-k-1)!} \tag{2.20}$$

for all $k\le(1-\delta)n$. Now by extensive use of Stirling's formula there is a constant $C>0$ such that

$$\begin{aligned}
&C_\delta n!k(n-k+1)!\frac{\binom{2n-2k}{n-k}}{(2n-k-1)!}\\
&\quad\le C_\delta Ck\sqrt{n}\left(\frac{n}{e}\right)^n\sqrt{n-k}\left(\frac{n-k}{e}\right)^{n-k}(n-k+1)\\
&\qquad\times\frac{(4^{n-k}/\sqrt{n-k})(2n-k)}{\sqrt{2n-k}((2n-k)/e)^{2n-k}}\\
&\quad= C_\delta Ck(n-k+1)\sqrt{n(2n-k)}2^{-k}\\
&\qquad\times\left(\left(1-\frac{k}{n}\right)^{n/k-1}\left(1-\frac{k}{2n}\right)^{-2n/k+1}\right)^k,
\end{aligned}$$

where

$$\begin{aligned}
\left(1-\frac{k}{n}\right)^{n/k-1}\left(1-\frac{k}{2n}\right)^{-2n/k+1} &\le \left(1-\frac{k}{2n}\right)^{2n/k-2}\left(1-\frac{k}{2n}\right)^{-2n/k+1}\\
&= \left(1-\frac{k}{2n}\right)^{-1}\\
&\le \left(1-\frac{1-\delta}{2}\right)^{-1}\\
&= \frac{2}{1+\delta}.
\end{aligned}$$

This means that, for all $\delta > 0$, there exists a constant C'_δ such that, for $k \le (1-\delta)n$ and sufficiently large n, we have

$$\begin{aligned} &n!T(n,k)\frac{\binom{2n-2k}{n-k}(1-\alpha)^{2n-k-1}}{(2n-k-1)!} \\ &\qquad \le C'_\delta n^2(1-\alpha)^{2n}k(1+\delta)^{-k}(1-\alpha)^{-k}. \end{aligned} \tag{2.21}$$

Since $\sum k(1+\delta)^{-k}(1-\alpha)^{-k}$ converges for sufficiently small α we have shown that $S_1 = O(n^2(1-\alpha)^{2n})$. Furthermore, if we assume that n is sufficiently large so that $(1+\delta)(1-\alpha) \ge (1+\frac{\delta}{2})$, then as the terms in the sum

$$\sum_{k=1}^{(1-\delta)n} \frac{1}{n^2(1-\alpha)^{2n}} n!T(n,k)\frac{\binom{2n-2k}{n-k}(1-\alpha)^{2n-k-1}}{(2n-k-1)!} \tag{2.22}$$

are dominated by the terms in

$$\sum_{k=1}^{\infty} C'_\delta k\left(1+\frac{\delta}{2}\right)^{-k}, \tag{2.23}$$

which converges, it follows by dominated convergence together with Proposition 2.12 that

$$\sum_{k=1}^{(1-\delta)n} \frac{1}{n^2(1-\alpha)^{2n}} n!T(n,k)\frac{\binom{2n-2k}{n-k}(1-\alpha)^{2n-k-1}}{(2n-k-1)!} \longrightarrow \sum_{k=1}^{\infty} k2^{1-k} = 4$$

as $n \to \infty$. □

PROPOSITION 2.14. *For sufficiently small $\delta > 0$ and $\alpha = o(1)$, we have $S_2 = O(n(1-\alpha)^n)$.*

PROOF. Using Proposition 2.11 there is a constant C such that this sum is bounded by

$$\begin{aligned} \sum_{l=0}^{\delta n} n!T(n,n-l)\frac{\binom{2l}{l}(1-\alpha)^{n+l-1}}{(n+l-1)!} &\le C\sum_{l=0}^{\delta n} n!(l+1)\left(\frac{n+2l}{5}\right)^l \frac{\binom{2l}{l}(1-\alpha)^{n+l-1}}{(n+l-1)!} \\ &\le C(1-\alpha)^{n-1}\sum_{l=0}^{\delta n} n^{1-l}(l+1)\left(\frac{n+2l}{5}\right)^l 4^l \\ &\le Cn(1-\alpha)^{n-1}\sum_{l=0}^{\infty}(l+1)\left(\frac{4(1+2\delta)}{5}\right)^l, \end{aligned}$$

where the last sum clearly converges for sufficiently small δ. □

PROPOSITION 2.15. *Let X be the number of accessible paths in the α-HoC model where $\alpha = \frac{\ln n}{n} - \varepsilon_n$ where $n\varepsilon_n \to \infty$. Then*

$$\mathbb{E}[X^2] \sim 4n^2(1-\alpha)^{2n}. \tag{2.24}$$

PROOF. From Proposition 2.3 together with Propositions 2.13 and 2.14 we know that

$$\mathbb{E}[X^2] \le (4+o(1))n^2(1-\alpha)^{2n} + O\big(n(1-\alpha)^n\big), \tag{2.25}$$

where one can show that $n(1-\alpha)^n = o(n^2(1-\alpha)^{2n})$, provided $n\varepsilon_n \to \infty$.

To derive a tight lower bound for $\mathbb{E}[X^2]$, consider the sum of $\mathbb{E}[X_i X_j]$ over all pairs of paths whose number of common interior nodes, $k-1$, is at most $\frac{n}{2}-1$ and that diverge at most once. Expressed in terms of components of permutations, for a fixed i and k, the number of paths j that satisfy this equals the number of permutations with k components, where all but one component contains exactly one element. This can clearly be done in $kT(n-k+1,1) \sim k(n-k+1)!$ ways.

By Proposition 2.3 this yields

$$\mathbb{E}[X^2] \ge \sum_{k=1}^{n/2} n!kT(n-k+1,1)\frac{\binom{2n-2k}{n-k}(1-\alpha)^{2n-k-1}}{(2n-k-1)!}. \tag{2.26}$$

Proceeding in a manner similar to the proof of Proposition 2.13, we get that

$$\sum_{k=1}^{n/2} n!kT(n-k+1,1)\frac{\binom{2n-2k}{n-k}(1-\alpha)^{2n-k-1}}{(2n-k-1)!} \sim 4n^2(1-\alpha)^{2n}, \tag{2.27}$$

which completes the proof. □

From this proof we can observe that almost all of the contributions to $\mathbb{E}[X^2]$ come from pairs of paths we considered in the lower bound. This implies the following:

COROLLARY 2.16. *Assume $\alpha = \frac{\ln n}{n} - \varepsilon_n$ where $n\varepsilon_n \to \infty$. For any $0 < \delta < 1$, the contribution to $\mathbb{E}[X^2]$ from all pairs of paths that either share more than $(1-\delta)n$ common nodes or that diverge more than once is $o(n^2(1-\alpha)^{2n})$.*

2.4. *Proof of Theorem 1.2.* Let X as above denote the number of accessible paths in α-HoC, where $\alpha = \frac{\ln n}{n} - \varepsilon_n$, $0 \le \varepsilon_n \le \frac{\ln n}{n}$ and $n\varepsilon_n \to \infty$. Applying Lemma 2.2 to X and using the expressions for $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ from Propositions 2.1 and 2.15, respectively, yields the lower bound

$$\liminf_{n\to\infty} P\left(n, \frac{\ln n}{n} - \varepsilon_n\right) \ge \frac{1}{4}. \tag{2.28}$$

In this subsection, we will prove that this probability can be "bootstrapped" up to 1, proving the remaining part of Theorem 1.2.

LEMMA 2.17. *Let $0 \le a \le 1-b \le 1$, and let $f : \mathbb{Q}_n \to \mathbb{R}$ be a fitness function whose values are generated independently according to*

$$f(v) = \begin{cases} a, & \text{if } v = \mathbf{v}^0, \\ 1-b, & \text{if } v = \mathbf{v}^1, \\ \sim \mathrm{U}(0,1), & \text{otherwise.} \end{cases} \tag{2.29}$$

Then the probability of accessible paths with respect to f equals $P(n, a+b)$.

PROOF. Define the function $g : \mathbb{Q}_n \to \mathbb{R}$ by setting $g(v) = f(v) + b$ if $f(v) \le 1-b$ and $g(v) = f(v) - 1 + b$ otherwise. Then $g(\mathbf{v}^0) = a+b$, $g(\mathbf{v}^1) = 1$ and $g(v) \sim \mathrm{U}(0,1)$ independently for all other v, so g is distributed as in α-HoC with $\alpha = a+b$. As this transformation only constitutes a translation for any node on an accessible path, we see that a path is accessible with respect to f if and only if it is so with respect to g. □

PROPOSITION 2.18. *Assume there is a positive constant C such that $\liminf_{n\to\infty} P(n, \frac{\ln n}{n} - \varepsilon_n) \ge C$ whenever $0 \le \varepsilon_n \le \frac{\ln n}{n}$ is a sequence satisfying $n\varepsilon_n \to \infty$. Then, the same inequality holds if C is replaced by $1-(1-C)(1-\frac{C}{2})$.*

PROOF. Let $\alpha = \frac{\ln n}{n} - \varepsilon_n$. We wish to pick four nodes, a_1, a_2, b_1, b_2, satisfying the following conditions:

(i) $d(a_1, \mathbf{v}^0) = d(a_2, \mathbf{v}^0) = 1$ and a_1, a_2 each has fitness in the range $[\alpha, \alpha + \varepsilon_n/3]$;
(ii) $d(b_1, \mathbf{v}^1) = d(b_2, \mathbf{v}^1) = 1$ and b_1, b_2 each has fitness at least $1 - \varepsilon_n/3$;
(iii) none of the four pairs (a_i, b_j) are antipodal (in the undirected hypercube).

By (i), the number of possibilities for each a_i is binomially distributed with parameters $\mathrm{Bin}(n, \varepsilon_n/3)$. Then, by (ii) and (iii), the number of options for each b_j is distributed as $\mathrm{Bin}(n-2, \varepsilon_n/3)$. Since $n\varepsilon_n/3 \to \infty$, it follows that it is possible to choose four nodes satisfying (i)–(iii) with probability $1 - o_n(1)$.

Condition on the fitness of all vertices v with $d(v, \mathbf{v}^0) = 1$ or $d(v, \mathbf{v}^1) = 1$. Let H_1 and H_2 be the induced subgraphs consisting of all nodes on paths from a_1 to b_1 and from a_2 to b_2, respectively, and let H_2' be the induced subgraph consisting of all nodes on paths between a_2 and b_2 that does not intersect H_1 in any vertex. Then H_1 and H_2 are isomorphic to $\mathbb{Q}_{n-2}$. Note that any accessible path from a_1 to b_1 or a_2 to b_2 can be extended to an accessible path from $\mathbf{v}^0$ to $\mathbf{v}^1$.

Let us denote the probability of accessible paths through the respective induced subgraphs by p_{H_1}, p_{H_2} and $p_{H_2'}$. By construction, H_1 and H_2' are vertex disjoint, so the events of accessible paths through the two subgraphs are independent. By Lemma 2.17, $p_{H_1} = P(n-2, f(a_1) + 1 - f(b_1)) \ge P(n-2, \alpha + \frac{2\varepsilon_n}{3})$. It is straightforward to show that this is still below the threshold, which implies that $p_{H_1} \ge C - o_n(1)$.

To estimate $p_{H_2'}$, we note that a path in H_2 from a_2 to b_2 is contained in H_2' if and only if it "flips the bit that is 1 in a_1 after that which is 0 in b_1." In the cases where there is an accessible path through H_2, let γ be chosen uniformly among all such paths. Then, by symmetry, we know that it flips the two bits corresponding to a_1 and b_1 in the allowed order, and is thus contained in H_2', with probability $\frac{1}{2}$. Hence $p_{H_2'} \geq \frac{1}{2} p_{H_2} = \frac{1}{2} p_{H_1}$.

As the events of accessible paths through H_1 and H_2' are independent, we get $P(n, \alpha) \geq 1 - (1 - p_{H_1})(1 - p_{H_2'}) - o_n(1) \geq 1 - (1 - C)(1 - \frac{C}{2}) - o_n(1)$ and the proposition follows. □

Now we complete the proof of Theorem 1.2. By equation (2.28) and repeated use of Proposition 2.18 we can construct a sequence $\{C_k\}_{k=0}^{\infty}$ such that $C_k \to 1$ and $\liminf_{n\to\infty} P(n, \alpha) \geq C_k$ for all k. Hence we must have $\liminf_{n\to\infty} P(n, \alpha) = 1$.

2.5. *Proof of Corollary* 1.4. Similarly to the proof of Theorem 1.2, that of Corollary 1.4 will use an alternative formulation of the α-HoC model. A key observation is that if one generates fitnesses according to α-HoC but then removes interior vertices independently with some probability δ, then this results in a model equivalent to α'-HoC for some $\alpha' > \alpha$. The intuition is that if α is far below the threshold $\frac{\ln n}{n}$, then not only is there an accessible path with probability $1 - o_n(1)$, but even if we remove a sufficient amount of vertices so that most paths become forbidden, we will still be below the threshold and so will still have accessible paths with probability $1 - o_n(1)$. This intuitively requires the original number of accessible paths to be large. Interestingly, this argument only requires the first equation in Theorem 1.2 even though the corollary itself is a stronger form of that statement.

This idea is formalized in the following lemmas:

LEMMA 2.19. *Let $\alpha, \delta \in [0, 1]$. Consider the fitness model that first assigns fitnesses as in α-HoC, but then independently removes each vertex in $\mathbb{Q}_n \setminus \{\mathbf{v}^0, \mathbf{v}^1\}$ with probability δ. Then the probability of accessible paths using only the remaining vertices is $P(n, 1 - (1 - \alpha)(1 - \delta))$.*

PROOF. Let $\alpha' = 1 - (1 - \alpha)(1 - \delta)$. We compare the model described above with α'-HoC.

Let us make the slight modification to α'-HoC and the above model that we additionally consider any vertex removed if it is less fit than $\mathbf{v}^0$. As no such node can be part of an accessible path, this will not change accessibility in either model. We see that these formulations are equivalent up to a translation and scaling, so they will have the same distribution of accessible paths. □

LEMMA 2.20. *Let Ω be a finite universal set, and let R be a random subset of Ω given by $\mathbb{P}(r \in R) = p_r$, these events being mutually independent over $r \in \Omega$. Let $\{A_i\}_{i\in I}$ be subsets of Ω, I a finite index set. Let B_i be the event $A_i \subseteq R$. Then*

$$\prod_{i\in I} \mathbb{P}(\bar{B}_i) \leq \mathbb{P}\Big(\bigwedge_{i\in I} \bar{B}_i\Big). \tag{2.30}$$

This inequality is commonly used as a lower bound in Janson's inequality. See, for instance, Theorem 8.1.1 in [2].

PROOF OF COROLLARY 1.4. The upper bound is simply Markov's inequality. We now turn to the lower bound. To simplify calculations we may, without loss of generality, assume that $w_n = o(n\varepsilon_n)$ and that $1 \leq w_n \leq e^{n\varepsilon_n}$ for all n.

Let $\delta_n = \varepsilon_n - \frac{\ln w_n}{n}$ and let Y denote the number of intact accessible paths using the same fitness function as for X but after removing each node except $\mathbf{v}^0$ and $\mathbf{v}^1$ independently with probability δ_n. By assumption, we know that $0 \leq \delta_n \leq \varepsilon_n \leq \frac{\ln n}{n}$, so δ_n is always a valid probability.

Using Lemma 2.19 we see that $\mathbb{P}(Y > 0) = P(n, \alpha_n')$ where $\alpha_n' = 1 - (1 - \alpha)(1 - \delta_n) = \frac{\ln n}{n} - \frac{o(1)+\ln w_n}{n}$. As $o(1) + \ln w_n \to \infty$ as $n \to \infty$ it follows from Theorem 1.2 that $\lim_{n\to\infty} \mathbb{P}(Y = 0) = 0$.

Condition on the set of accessible paths before removing vertices. Let I be the set of accessible paths, R the random set of nonremoved vertices and B_i the event that path $i \in I$ only consist of nonremoved vertices. Then we are in the setting of Lemma 2.20. As the probability that each accessible path remains intact is $(1 - \delta_n)^{n-1}$, averaging conditioned on X we get the inequality

$$\mathbb{P}(Y = 0 \mid X) \geq \big(1 - (1 - \delta_n)^{n-1}\big)^X. \tag{2.31}$$

But since $\lim_{n\to\infty} \mathbb{P}(Y = 0) = 0$ and $(1 - (1 - \delta_n)^{n-1})^X = e^{-(1+o(1))e^{-n\delta_n} X}$ it follows that $e^{-n\delta_n} X$ must tend to infinity in probability. To complete the proof we note that $e^{-n\delta_n} X = \frac{X}{e^{n\varepsilon_n}/w_n} \sim \frac{X}{\mathbb{E}[X]/w_n}$. $\square$

REMARK 2.21. Note that Proposition 2.15 implies that $\mathrm{Var}(X) \sim 3\mathbb{E}[X]^2$ for α in this regime, so no significant improvement on Corollary 1.4 can be made by a naive application of Chebyshev's inequality.

3. Results for the RMF model. Let $n \in \mathbb{N}$, and let $\varepsilon = \varepsilon_n$ be some strictly positive function. Consider the n-dimensional hypercube in which $\mathbf{v}^0$ and $\mathbf{v}^1$ are present, and where every other vertex is present with probability ε_n, independently of all other vertices. Let $Y = Y_{n,\varepsilon_n}$ denote the number of accessible paths from $\mathbf{v}^0$ to $\mathbf{v}^1$, where in this model a path is accessible if Hamming distance from $\mathbf{v}^0$ is strictly increasing and all vertices along the path are present. The following proposition may be well known, as it can be interpreted in the context of site percolation on the directed hypercube. However, we were not able to locate a suitable reference.

PROPOSITION 3.1. (i) $\mathbb{E}[Y] = n! \cdot \varepsilon_n^{n-1}$.

(ii) *Let* $n \to \infty$, *and suppose that* $n\varepsilon_n \to \infty$. *Then* $\mathrm{Var}(Y) = o(\mathbb{E}[Y]^2)$, *and hence*

$$Y \sim \mathbb{E}[Y] \sim \frac{\sqrt{2\pi n}}{\varepsilon_n}\left(\frac{n\varepsilon_n}{e}\right)^n. \tag{3.1}$$

PROOF. There are $n!$ possible paths in the n-hypercube. Each path contains $n-1$ interior vertices, each of which is present with probability ε_n. This proves (i). Set $\mu = \mu_n := n!\varepsilon_n^{n-1}$. Now suppose $n\varepsilon_n \to \infty$. Let Y_i be the indicator of the event that the ith increasing path is accessible, where the paths have been ordered in any way. Fix any path i_0. Then, by a standard second moment estimate (see Section 2),

$$\mathrm{Var}(Y) \le \mu + n! \cdot \sum_{j \sim i_0} \mathbb{E}(Y_{i_0} Y_j), \tag{3.2}$$

where the sum is taken over all paths j which intersect the path i_0 in at least one interior vertex. Let k be the number of intersection points. This leaves $T(n, k+1)$ possibilities for the path j. The paths i_0 and j contain a total of $2n-2-k$ different interior vertices; hence the probability of both being present is ε_n^{2n-2-k}. Hence

$$\mathrm{Var}(Y) \le \mu + n! \cdot \sum_{k=2}^{n} T(n,k)\varepsilon_n^{2n-1-k} \le \mu + \mu^2 \cdot \sum_{k=2}^{n} \frac{T(n,k)}{n!\varepsilon_n^{k-1}}. \tag{3.3}$$

Hence since $\mu \to \infty$ when $n\varepsilon_n \to \infty$, it suffices to show that

$$\sum_{k=2}^{n} \frac{T(n,k)}{n!\varepsilon_n^{k-1}} = o(1). \tag{3.4}$$

We now follow the same strategy as in Section 2, but the analysis here is much simpler. Let $\delta \in (0,1)$. We divide the sum in (3.4) into two parts, one for $k \le (1-\delta)n$ and the other for $k > (1-\delta)n$. From Proposition 2.9 and Lebesgue's dominated convergence theorem, it follows easily that, for any $\delta > 0$, the sum over terms $k \le (1-\delta)n$ is bounded by $(1 + o_n(1)) \sum_{k=2}^{\infty} \frac{k}{(n\varepsilon_n)^{k-1}} = O(\frac{1}{n\varepsilon_n}) = o(1)$, provided $n\varepsilon_n \to \infty$. Similarly, from Proposition 2.11 it follows that the sum over terms $k > (1-\delta)n$ is bounded by

$$\frac{c}{\mu} \sum_{l=0}^{\delta n} (l+1)\left(\frac{1+2\delta}{5} \cdot n\varepsilon_n\right)^l, \tag{3.5}$$

where c is an absolute constant. Since $n\varepsilon_n \to \infty$, the sum in (3.5) is bounded by $1 + o(1)$ times the last term, and hence is $O((n\varepsilon_n)^{\delta n})$, which is in turn $o(\mu)$. This proves (3.4) and completes the proof of the proposition. □

We now turn to the RMF model and prove Theorem 1.5.

We shall abuse notation and also use η to denote the p.d.f. of the probability distribution under consideration. So suppose η has connected support and is continuous there. Let $\delta > 0$ be given. Then there exists a bounded, closed interval $I = I_\delta \subseteq \mathrm{Supp}(\eta)$ such that $\int_{I_\delta} \eta(x)\,dx > 1 - \delta$. The quantity $c_{\eta,\delta} = \min_{x \in I_\delta} \eta(x)$ exists, is nonzero and, obviously, depends only on η and δ. Now let $n \in \mathbb{N}$ and $\theta = \theta_n > 0$ be given. Without loss of generality, we may assume that the interval I_δ has length $l(I_\delta) > \theta_n/2$ (in fact any multiple $c\theta_n$, where $0 < c < 1$, would do in the argument that follows). By definition of I_δ, with probability at least $(1-\delta)^2$ each of $\eta(\mathbf{v}^0)$ and $\eta(\mathbf{v}^1)$ lie in I_δ. Let X_{δ,n,θ_n} be the number of accessible paths in the n-hypercube, where fitnesses are assigned as in (1.2), and conditioning on the fact that both $\eta(\mathbf{v}^0)$ and $\eta(\mathbf{v}^1)$ lie in I_δ. We claim that, if n is sufficiently large, then X_{δ,n,θ_n} stochastically dominates the random variable Y_{n,ε_n} in Proposition 3.1, where $\varepsilon_n = c_{\eta,\delta} \cdot \frac{\theta_n}{2}$.

To see this, first note that, as long as $l(I_\delta) > \theta_n/2$ then, for any point $x \in I_\delta$, there will be an interval I_x of length at least $\theta_n/2$, which contains x and lies entirely within I_δ. By assumption, any such interval captures at least $c_{\eta,\delta} \cdot \frac{\theta_n}{2}$ of the distribution η. For any adjacent pair (v, v') of vertices in the hypercube such that $d(v', \mathbf{v}^0) = d(v, \mathbf{v}^0) + 1$, if $\eta(v') > \eta(v) - \theta_n$, then v' is accessible from v. Assuming $\eta(\mathbf{v}^0) \in I_\delta$, it follows that we can choose, for each layer i in the hypercube, an interval $I_i \subseteq I_\delta$ of length $\theta_n/2$ such that any path

$$\mathbf{v}^0 \to v_1 \to v_2 \to \cdots \to v_{n-1} \tag{3.6}$$

for which $\eta(v_i) \in I_i$ for all $i = 1, \ldots, n-1$, is accessible. If n is sufficiently large, we can also ensure that the interval I_{n-1} contains $\eta(\mathbf{v}^1)$, so that any viable path (3.6) can definitely be continued to $\mathbf{v}^1$. The stochastic domination of Y_{n,ε_n} by X_{δ,n,θ_n} now follows. Then one just needs to apply Proposition 3.1 and Theorem 1.5 follows immediately.

REMARK 3.2. Suppose $\mathrm{Supp}(\eta)$ is also bounded and that θ is a constant, independent of n. Let

$$C_{\eta,\theta} := \min_{l(I)=\theta/2, I \subseteq \overline{\mathrm{Supp}(\eta)}} \int_I \eta(x)\,dx, \tag{3.7}$$

where I denotes a closed interval. Then this minimum exists and is nonzero. It follows from Proposition 3.1 and the argument above that the number $X = X(n)$ of accessible paths in this case satisfies

$$X \gtrsim n! \cdot C_{\eta,\theta}^{n-1}. \tag{3.8}$$

The point is that $C_{\eta,\theta} \in (0, 1]$ is a constant depending only on η and θ.

Acknowledgements. We thank Joachim Krug for making us aware of the problems studied here, and both he and Stefan Nowak for helpful discussions. We thank both referees for their very careful reading of the manuscript.

REFERENCES

[1] AITA, T., UCHIYAMA, H., INAOKA, T., NAKAJIMA, M., KOKUBO, T. and HUSIMI, Y. (2000). Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape: Application to prolyl endopeptidase and thermolysin. *Biopolymers* **54** 64–79.

[2] ALON, N. and SPENCER, J. H. (2008). *The Probabilistic Method*, 3rd ed. Wiley, Hoboken, NJ. MR2437651

[3] CARNEIRO, M. and HARTL, D. L. (2010). Colloquium papers: Adaptive landscapes and protein evolution. *Proc. Natl. Acad. Sci. USA* **107 Suppl 1** 1747–1751.

[4] COMTET, L. (1972). Sur les coefficients de l'inverse de la série formelle $\sum n!t^n$. *C. R. Acad. Sci. Paris Sér. A–B* **275** A569–A572. MR0302457

[5] COMTET, L. (1974). *Advanced Combinatorics*: *The Art of Finite and Infinite Expansions*, enlarged ed. Reidel, Dordrecht. MR0460128

[6] FRANKE, J., KLÖZER, A., DE VISSER, J. A. G. M. and KRUG, J. (2011). Evolutionary accessibility of mutational pathways. *PLoS Comput. Biol.* **7** e1002134, 9. MR2845072

[7] FRANKE, J. and KRUG, J. (2012). Evolutionary accessibility in tunably rugged fitness landscapes. *J. Stat. Phys.* **148** 705–722.

[8] FRANKE, J., WERGEN, G. and KRUG, J. (2010). Records and sequences of records from random variables with a linear trend. *J. Stat. Mech. Theory Exp.* **10** P10013, 21. MR2800498

[9] KAUFFMAN, S. and LEVIN, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *J. Theoret. Biol.* **128** 11–45. MR0907587

[10] KINGMAN, J. F. C. (1978). A simple model for the balance between selection and mutation. *J. Appl. Probab.* **15** 1–12. MR0465272

[11] KLÖZER, A. (2008). NK fitness landscapes. Diplomarbeit Universität zu Köln.

[12] MÉZARD, M., PARISI, G. and VIRASORO, M. A. (1987). *Spin Glass Theory and Beyond*. *World Scientific Lecture Notes in Physics* **9**. World Scientific, Teaneck, NJ. MR1026102

[13] THE ONLINE ENCYCLOPEDIA OF INTEGER SEQUENCES. Sequence #*A*003319. Available at http://oeis.org/A003319.

[14] THE ONLINE ENCYCLOPEDIA OF INTEGER SEQUENCES. Sequence #*A*059438. Available at http://oeis.org/A059438.

[15] WEINREICH, D. M., DELANEY, N. F., DEPRISTO, M. A. and HARTL, D. M. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312** 111–114.

[16] WEINREICH, D. M., WATSON, R. A. and CHAO, L. (2005). Perspective: Sign epistasis and genetic constraints on evolutionary trajectories. *Evolution* **59** 1165–1174.

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY AND
UNIVERSITY OF GOTHENBURG
41296 GOTHENBURG
SWEDEN
E-MAIL: hegarty@chalmers.se
andemar@chalmers.se